Statistical Learning STA4042

REGRESSION VS. CLASSIFICATION

SCHOOL OF DATA SCIENCE

Examples of Classification

Statistical Learning STA4042

REGRESSION VS. CLASSIFICATION

SCHOOL OF DATA SCIENCE

Examples of Classification

- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions.
Which of the three conditions does the individual has?

Statistical Learning STA4042

REGRESSION VS. CLASSIFICATION

SCHOOL OF DATA SCIENCE

Examples of Classification

- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions.
Which of the three conditions does the individual has?
- An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.

Statistical Learning STA4042

REGRESSION VS. CLASSIFICATION

SCHOOL OF DATA SCIENCE

Examples of Classification

- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions.
Which of the three conditions does the individual has?
- An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
- On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Statistical Learning STA4042

REGRESSION VS. CLASSIFICATION

SCHOOL OF DATA SCIENCE

Examples of Classification

- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions.
Which of the three conditions does the individual has?
- An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
- On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Same as for regression:
Training data set: $(x_1, y_1), ..., (x_n, y_n)$

- $x_1, \ldots, x_n \sim X$: predictors "quatitative" or "continuous")
- $y_1, \ldots, y_n \sim Y$: classes ("qualitative" or "discrete")

Statistical Learning STA4042

REGRESSION VS. CLASSIFICATION

SCHOOL OF DATA SCIENCE

Examples of Classification

- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions.
Which of the three conditions does the individual has?
- An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
- On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Same as for regression:
Training data set: $(x_1, y_1), ..., (x_n, y_n)$

Sometimes qualitative

- $x_1, \ldots, x_n \sim X$: predictors "quatitative" or "continuous")
- $y_1, \ldots, y_n \sim Y$: classes ("qualitative" or "discrete")

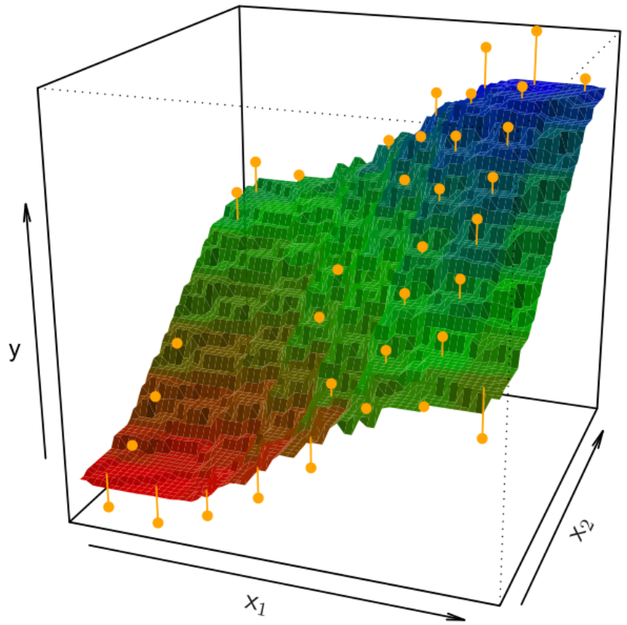
From regression to classification with KNN.

Training dataset:
$D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ given

New data test $x \in \mathbb{R}^p$:

$i_1, \ldots, i_k$ s.t.: $x_{i_1}, \ldots, x_{i_k}$ are the k closest neighbors of x:

$$\forall i \in [n]: \quad \|x - x_i\| \geq \sup_{j \in [k]} \|x - x_{i_k}\|.$$



- For regression:

$$\hat{f}_D(x) \equiv \frac{1}{k} \sum_{j=1}^{k} y_{i_j}.$$

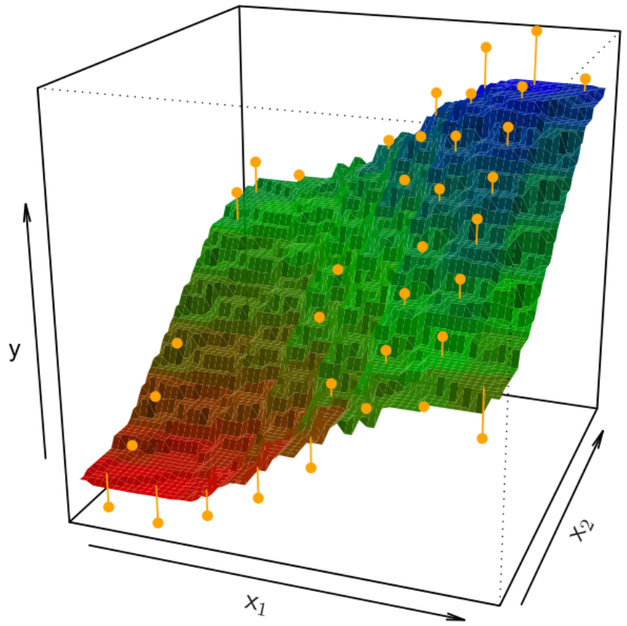
From regression to classification with KNN.

Training dataset:
$D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ given

New data test $x \in \mathbb{R}^p$:

$i_1, \ldots, i_k$ s.t.: $x_{i_1}, \ldots, x_{i_k}$ are the k closest neighbors of x:

$$\forall i \in [n] : \quad \|x - x_i\| \geq \sup_{j \in [k]} \|x - x_{i_k}\|.$$

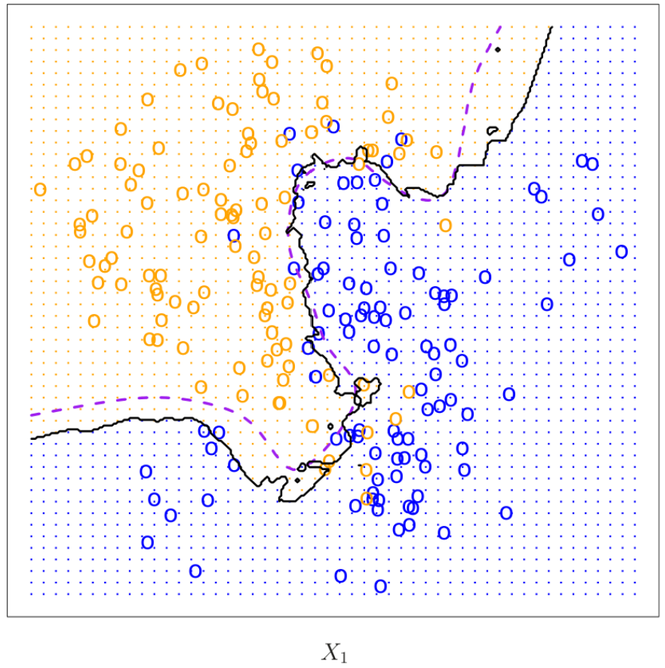


- For regression:

$$\hat{f}_D(x) \equiv \frac{1}{k} \sum_{j=1}^{k} y_{i_j}.$$

- For Classification $\hat{f}_D(x) = l$ where:

$$\forall h \in [k] : \quad \#\{y_{i_j} = l\} \geq \#\{y_{i_j} = h\}$$



From regression to classification with KNN.

k-nearest neighbors method:

Training dataset:
$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ given

New data test $x \in \mathbb{R}^p$:

$i_1, \dots, i_k$ s.t.: $x_{i_1}, \dots, x_{i_k}$ are the k closest neighbors of x:

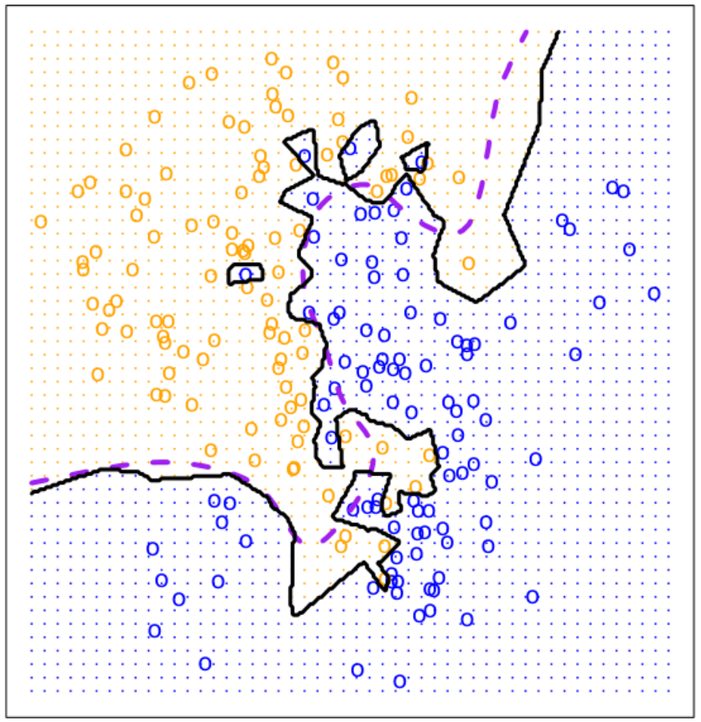
$$\forall i \in [n]: \quad \|x - x_i\| \geq \sup_{j \in [k]} \|x - x_{i_k}\|.$$

$\hat{f}_D(x) = l$ where $\forall h \in [k]$:

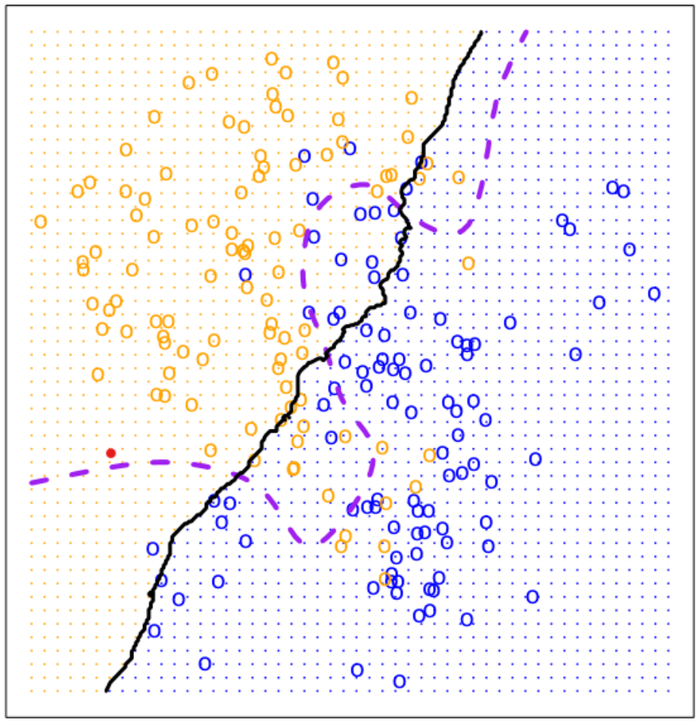
$$\#\{y_{i_j} = l\} \geq \#\{y_{i_j} = h\}$$

Trade-off on the number of neighbors k:

$k = 1$



$k = 100$



Why not linear regression?

Choose $$Y = \begin{cases} 1 & \text{if drug overdose} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if stroke} \end{cases}$$

Why not linear regression?

Choose $Y = \begin{cases} 1 & \text{if drug overdose} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if stroke} \end{cases}$ or $Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if drug overdose} \end{cases}$?

Why not linear regression?

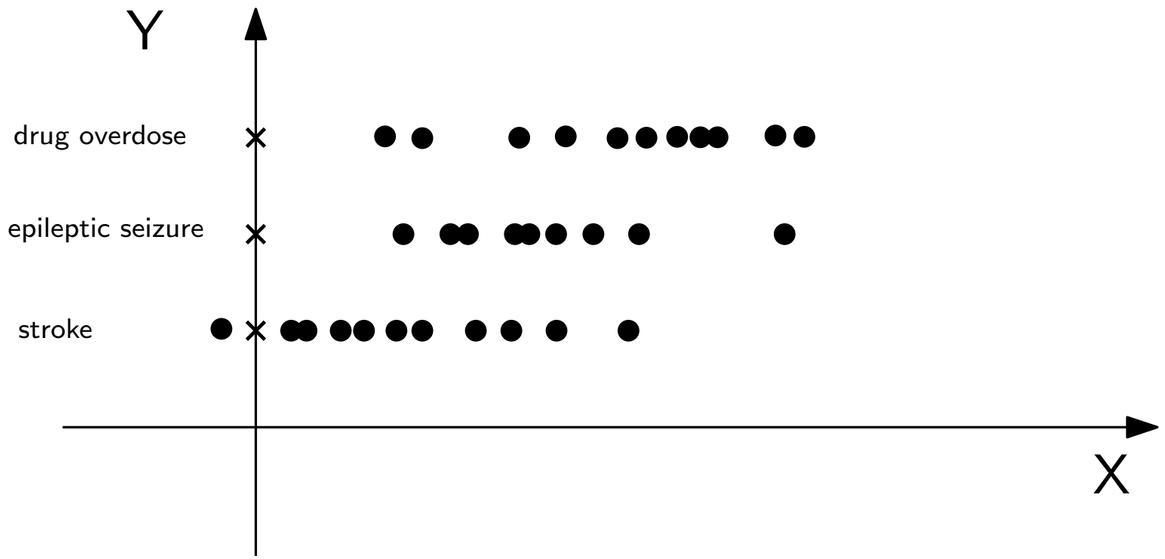
Choose $Y = \begin{cases} 1 & \text{if drug overdose} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if stroke} \end{cases}$ or $Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if drug overdose} \end{cases}$?

Why would we put one value in between ?

Why not linear regression?

Choose $Y = \begin{cases} 1 & \text{if drug overdose} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if stroke} \end{cases}$ or $Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if drug overdose} \end{cases}$?

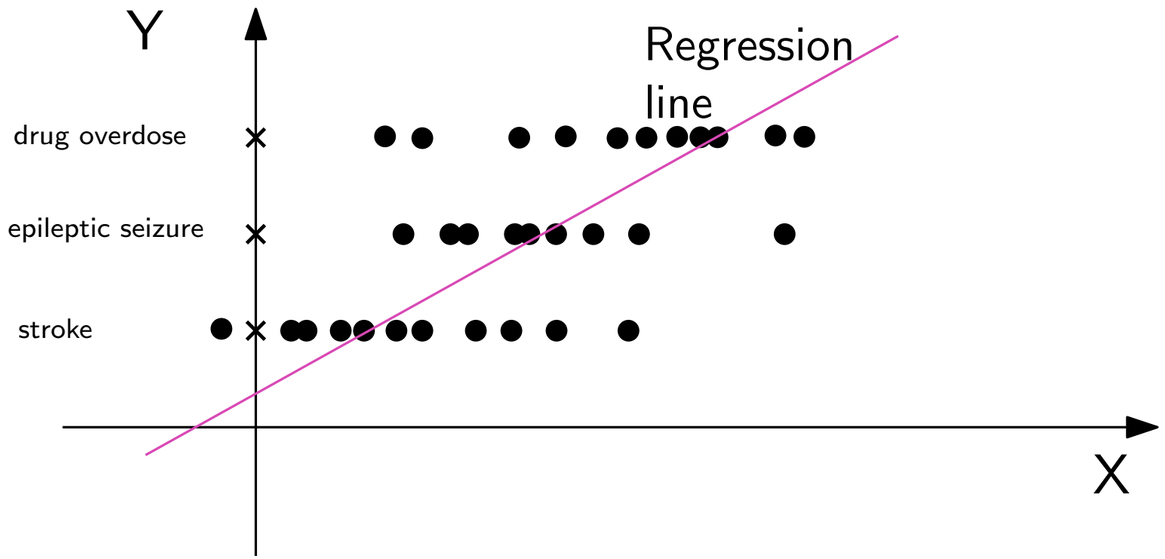
Why would we put one value in between ?



Why not linear regression?

Choose $Y = \begin{cases} 1 & \text{if drug overdose} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if stroke} \end{cases}$ or $Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if drug overdose} \end{cases}$?

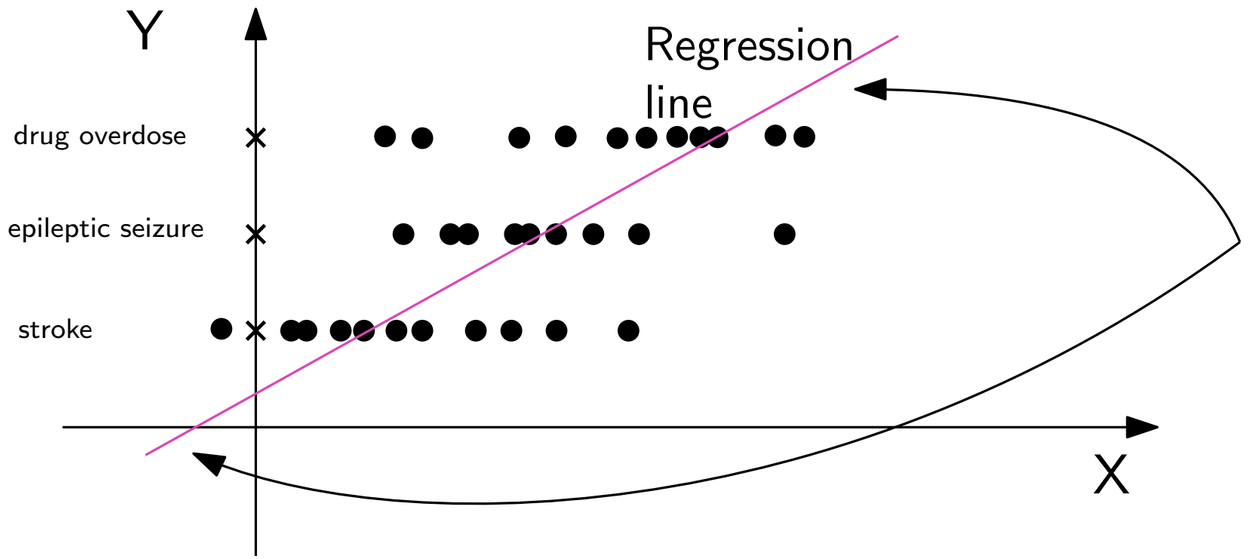
Why would we put one value in between ?



Why not linear regression?

Choose $Y = \begin{cases} 1 & \text{if drug overdose} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if stroke} \end{cases}$ or $Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if epileptic seizure} \\ 3 & \text{if drug overdose} \end{cases}$?

Why would we put one value in between ?



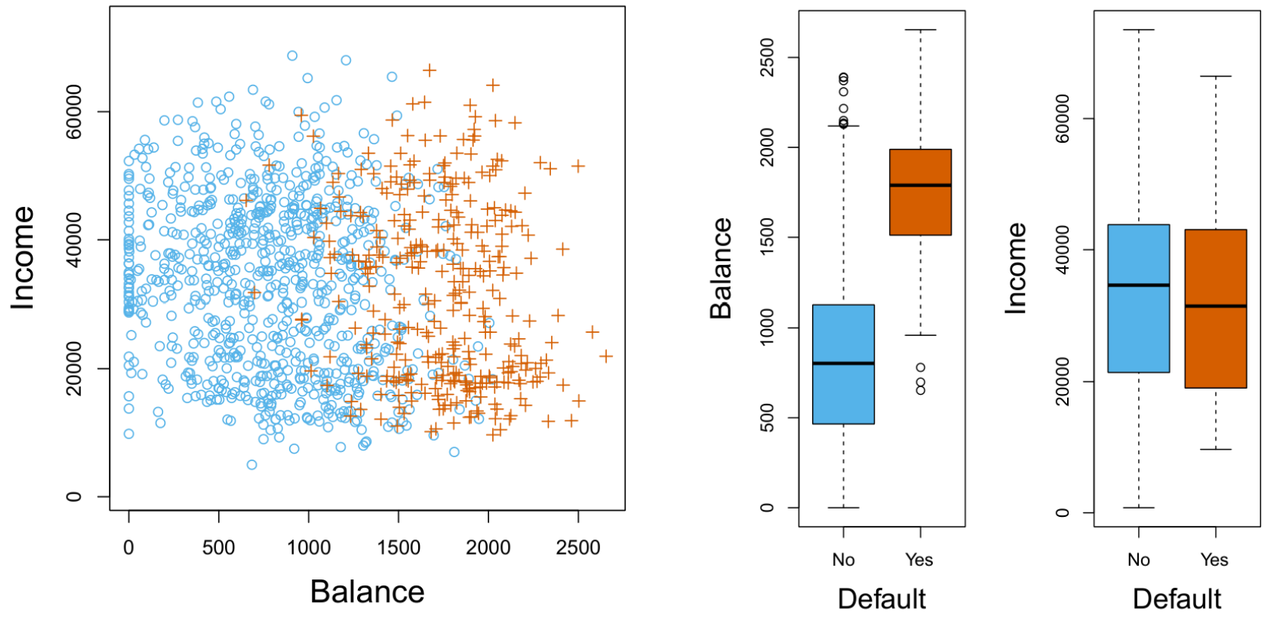
Negative and out of $[0, 1]$ predictions: not much sense!

Why not linear regression?

Back to two classes

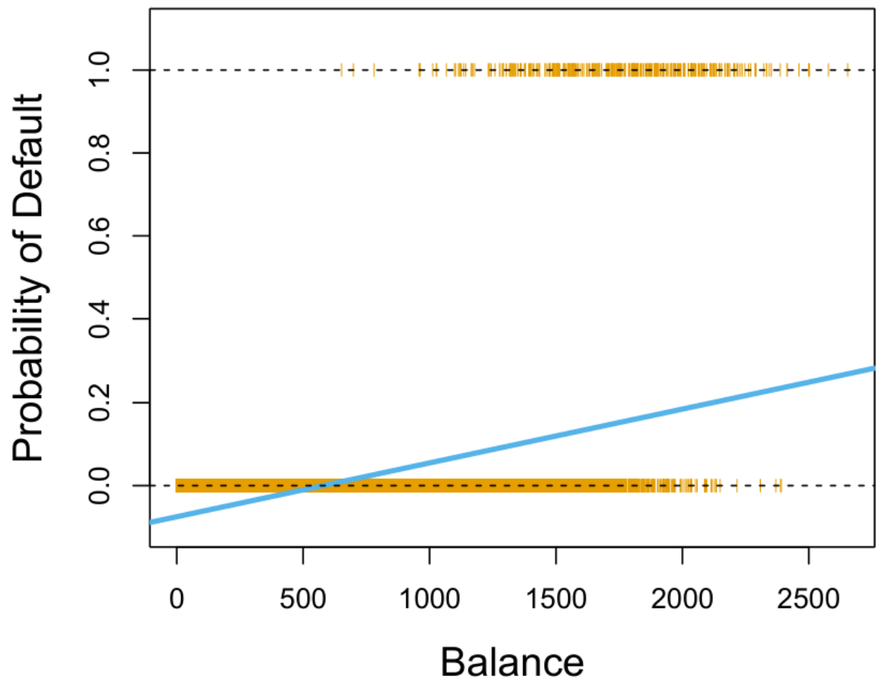
The annual incomes and monthly credit card balances of a number of individuals.

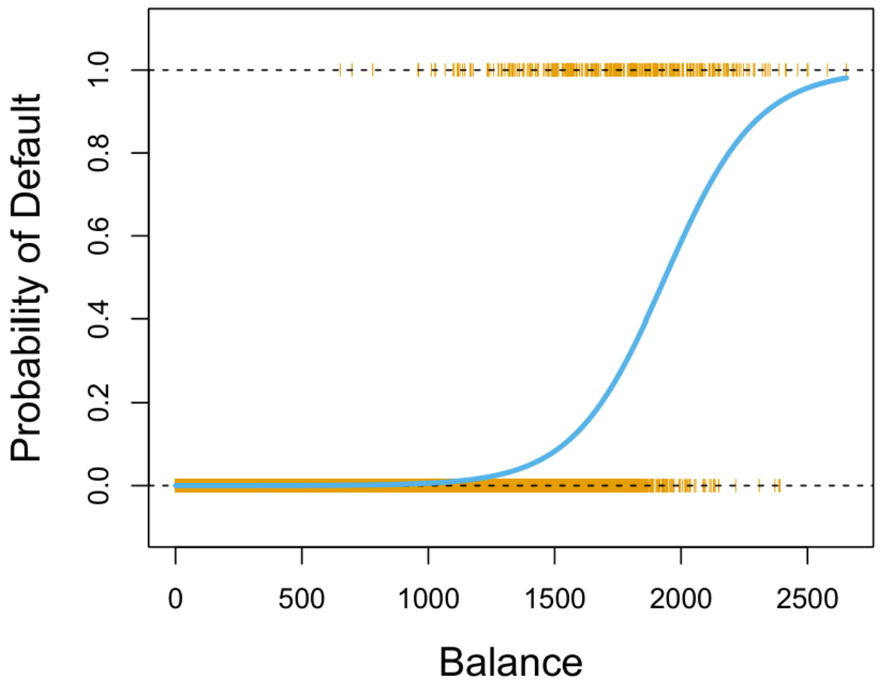
Orange: defaulted, **Blue**: did not default.



Why not linear regression?

Look for an output that stays close to the discrete values

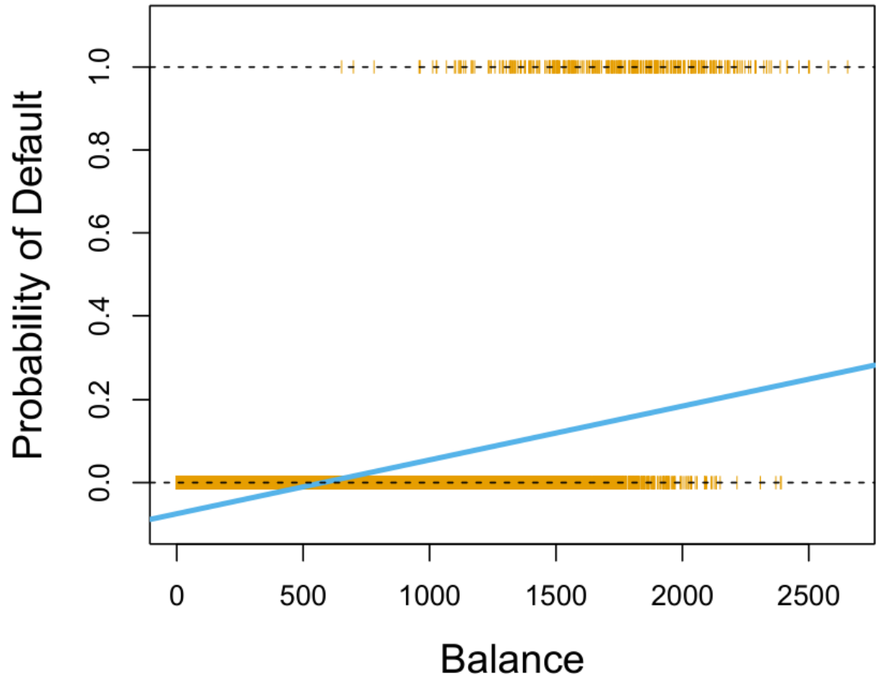




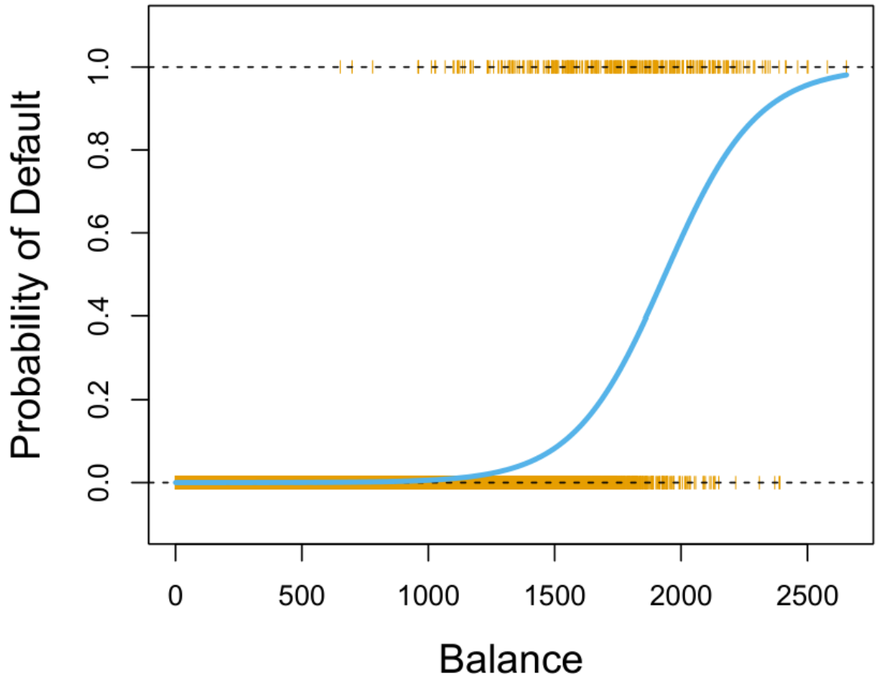
Linear regression

Why not linear regression?

Look for an output that stays close to the discrete values



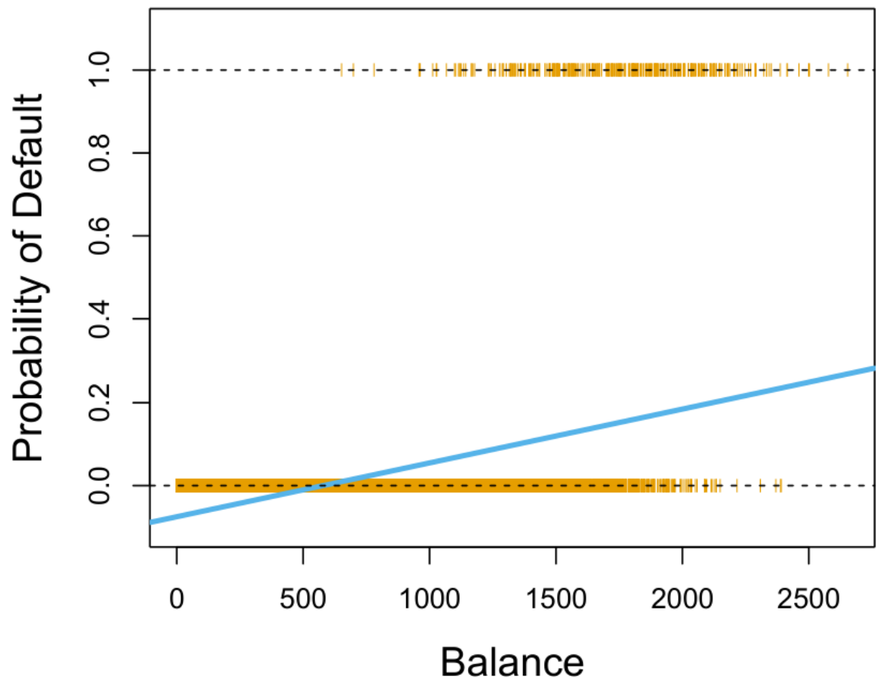
Linear regression



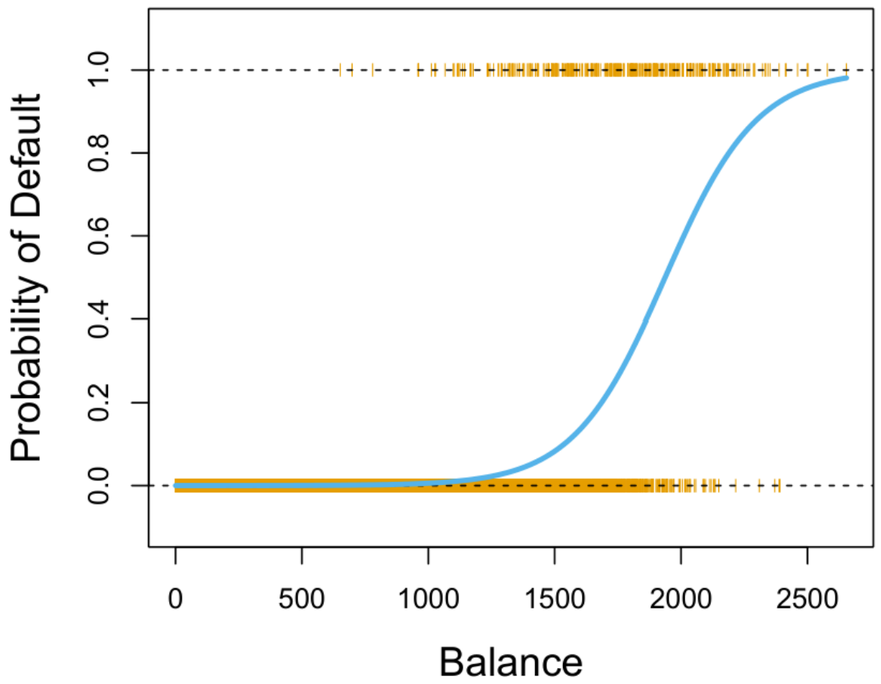
Logistic regression

Why not linear regression?

Look for an output that stays close to the discrete values



Linear regression

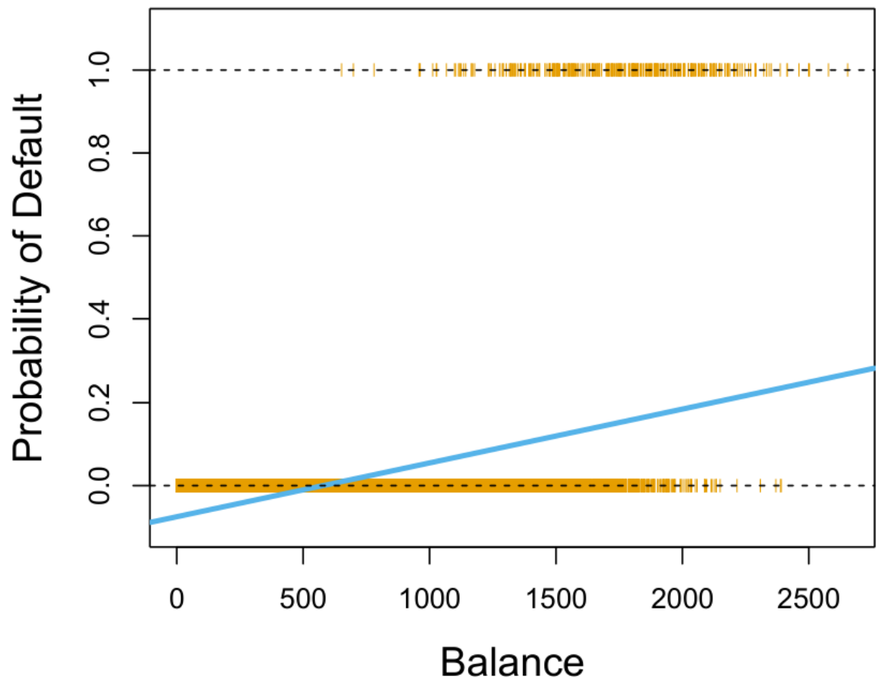


Logistic regression

$$\mathbb{P}(Y = 1) = \beta_0 + \beta_1 X$$

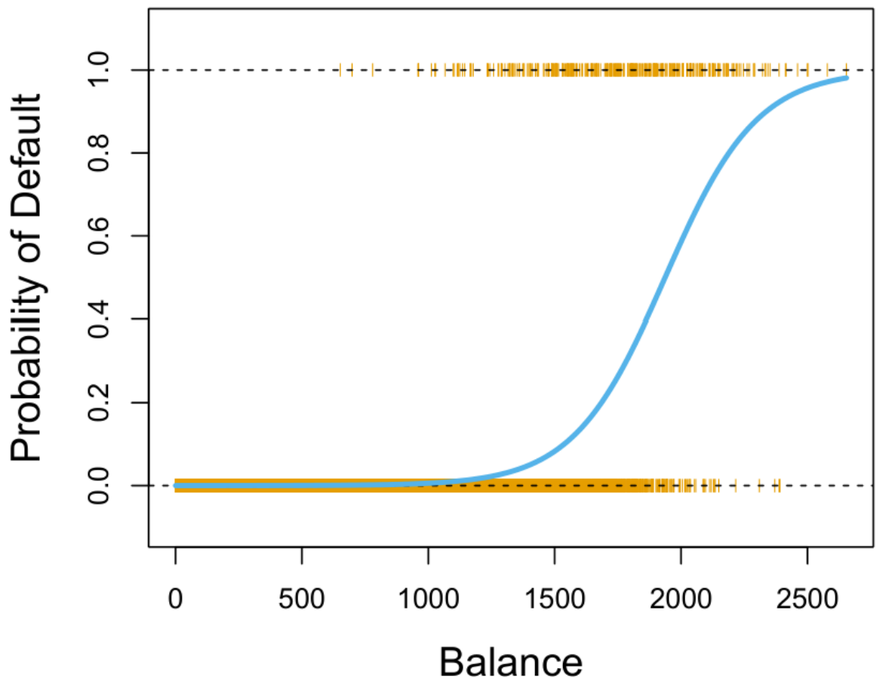
Why not linear regression?

Look for an output that stays close to the discrete values



Linear regression

$$\mathbb{P}(Y = 1) = \beta_0 + \beta_1 X$$



Logistic regression

$$\mathbb{P}(Y = 1) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \in [0, 1]$$

Why not linear regression?

When more than 2 classes?

Why not linear regression?

When more than 2 classes?

Need a different encoding:

$$\text{Ex: } Y = \begin{cases} (1,0,0) & \text{if drug overdose} \\ (0,1,0) & \text{if epileptic seizure} \\ (0,0,1) & \text{if stroke} \end{cases}$$

$$\mathbb{P}(Y = k) = \frac{e^{\beta_{k,0}+\beta_{k,1}X}}{\sum_{l=1}^{3} e^{\beta_{l,0}+\beta_{l,1}X}} \in [0,1]$$

Why not linear regression?

When more than 2 classes?

Need a different encoding:

$e_1, e_2, 2_3$: "one-hot vectors"

$$\text{Ex: } Y = \begin{cases} (1,0,0) & \text{if drug overdose} \\ (0,1,0) & \text{if epileptic seizure} \\ (0,0,1) & \text{if stroke} \end{cases}$$

$$\mathbb{P}(Y = k) = \frac{e^{\beta_{k,0} + \beta_{k,1} X}}{\sum_{l=1}^{3} e^{\beta_{l,0} + \beta_{l,1} X}} \in [0,1]$$

Why not linear regression?

When more than 2 classes?

Need a different encoding:

$e_1, e_2, 2_3$: "one-hot vectors"

All at the same distance from one another

$$\text{Ex: } Y = \begin{cases} (1,0,0) & \text{if drug overdose} \\ (0,1,0) & \text{if epileptic seizure} \\ (0,0,1) & \text{if stroke} \end{cases}$$

$$\mathbb{P}(Y = k) = \frac{e^{\beta_{k,0} + \beta_{k,1} X}}{\sum_{l=1}^{3} e^{\beta_{l,0} + \beta_{l,1} X}} \in [0,1]$$

Why not linear regression?

When more than 2 classes?

Need a different encoding:

$e_1, e_2, 2_3$: "one-hot vectors"

All at the same distance from one another

Not relevant for our simple methods, useful for neural network training

$$\text{Ex: } Y = \begin{cases} (1,0,0) & \text{if drug overdose} \\ (0,1,0) & \text{if epileptic seizure} \\ (0,0,1) & \text{if stroke} \end{cases}$$

$$\mathbb{P}(Y = k) = \frac{e^{\beta_{k,0} + \beta_{k,1} X}}{\sum_{l=1}^{3} e^{\beta_{l,0} + \beta_{l,1} X}} \in [0,1]$$

Why not linear regression?

When more than 2 classes?

Need a different encoding:

$e_1, e_2, 2_3$: "one-hot vectors"

All at the same distance from one another

Not relevant for our simple methods, useful for neural network training

$$\text{Ex: } Y = \begin{cases} (1,0,0) & \text{if drug overdose} \\ (0,1,0) & \text{if epileptic seizure} \\ (0,0,1) & \text{if stroke} \end{cases}$$

$$\mathbb{P}(Y = k) = \frac{e^{\beta_{k,0} + \beta_{k,1} X}}{\sum_{l=1}^{3} e^{\beta_{l,0} + \beta_{l,1} X}} \in [0,1]$$

$$\mathbb{P}(Y = 1) + \mathbb{P}(Y = 2) + \mathbb{P}(Y = 3) = 1$$

More details on the setting

Want to understand the relation between the predictors X and the class Y.

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$ or $Y \in \{e_1, \dots, e_k\} \subset \mathbb{R}^k$, if $k > 1$: "multinomial" or "multiclass" classification

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$ or $Y \in \{e_1, \dots, e_k\} \subset \mathbb{R}^k$, if $k > 1$: "multinomial" or "multiclass" classification

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$ or $Y \in \{e_1, \dots, e_k\} \subset \mathbb{R}^k$, if $k > 1$: "multinomial" or "multiclass" classification

Dataset or "observations": $(x_1, y_1), \dots, (x_n, y_n)$ $= n$ drawings of (X, Y)

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$, if $k > 1$: "multinomial" or "multiclass" classification

Dataset or "observations": $(x_1, y_1), \dots, (x_n, y_n)$ $= n$ drawings of (X, Y)

$\forall i \in [n] : (x_i, y_i) \in \mathbb{R}^p \times \{1, \dots, k\}$

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$, if $k > 1$: "multinomial" or "multiclass" classification

Dataset or "observations": $(x_1, y_1), \dots, (x_n, y_n)$ $= n$ drawings of (X, Y)

$\forall i \in [n] : (x_i, y_i) \in \mathbb{R}^p \times \{1, \dots, k\}$

NB: Often use notation $X = (X_1, \dots, X_p) \in \mathbb{R}^p$.

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$, if $k > 1$: "multinomial" or "multiclass" classification

Dataset or "observations": $(x_1, y_1), \dots, (x_n, y_n)$ $= n$ drawings of (X, Y)

$\forall i \in [n] : (x_i, y_i) \in \mathbb{R}^p \times \{1, \dots, k\}$

NB: Often use notation $X = (X_1, \dots, X_p) \in \mathbb{R}^p$.

Predictors

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$, if $k > 1$: "multinomial" or "multiclass" classification

Dataset or "observations": $(x_1, y_1), \dots, (x_n, y_n)$ $= n$ drawings of (X, Y)

$\forall i \in [n] : (x_i, y_i) \in \mathbb{R}^p \times \{1, \dots, k\}$

NB: Often use notation $X = (X_1, \dots, X_p) \in \mathbb{R}^p$.

Predictors

$x_i = (x_1^{(i)}, \dots, x_p^{(i)}) \in \mathbb{R}^p$

More details on the setting

Want to understand the relation between the predictors X and the class Y.

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$ or $Y \in \{e_1, \dots, e_k\} \subset \mathbb{R}^k$, if $k > 1$: "multinomial" or "multiclass" classification

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$ or $Y \in \{e_1, \dots, e_k\} \subset \mathbb{R}^k$, if $k > 1$: "multinomial" or "multiclass" classification

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$ or $Y \in \{e_1, \dots, e_k\} \subset \mathbb{R}^k$, if $k > 1$: "multinomial" or "multiclass" classification

Dataset or "observations": $(x_1, y_1), \dots, (x_n, y_n)$ $= n$ drawings of (X, Y)

More details on the setting

Want to understand the relation between the predictors X and the class Y.

Random vectors

$X \in \mathbb{R}^p$, if $p > 1$: "multiple" or "multivariate" classification

$Y \in \{1, \dots, k\} \subset \mathbb{R}$ or $Y \in \{e_1, \dots, e_k\} \subset \mathbb{R}^k$, if $k > 1$: "multinomial" or "multiclass" classification

Dataset or "observations": $(x_1, y_1), \dots, (x_n, y_n)$ $= n$ drawings of (X, Y)

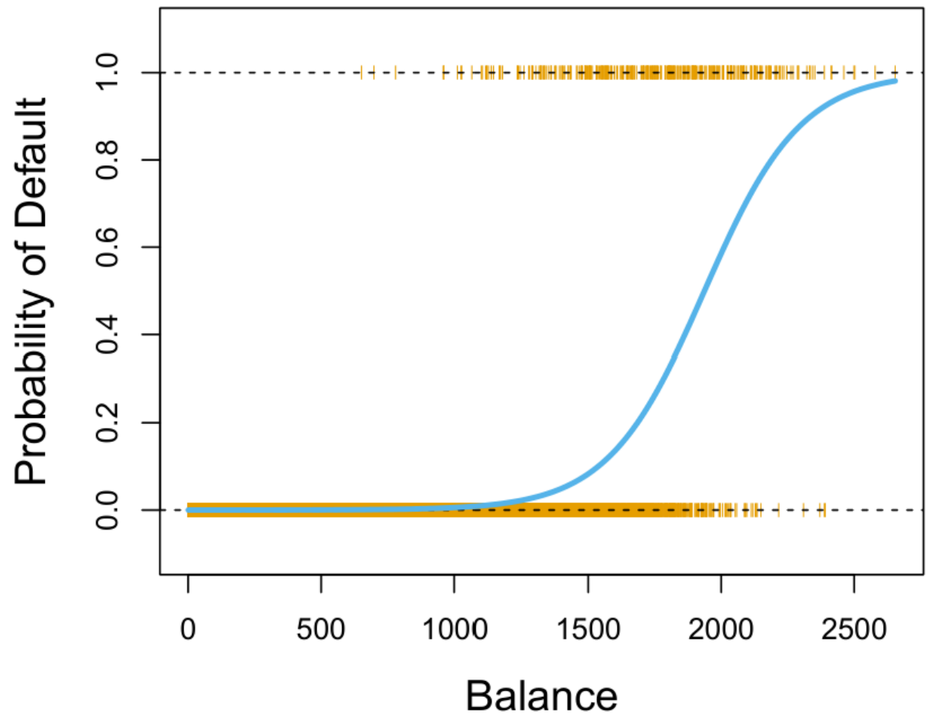
Goal: New observation $x \sim X \rightarrow$ good estimation of class y ?

Logistic regression

Model: $\mathbb{P}(Y=1) = s(\beta_0 + \beta_1 X)$ with $s : t \mapsto \frac{e^t}{1+e^t}$ **(sigmoid)**

Minimize negative loglikelihood:

$$NL(\beta) = \sum_{y_i=0} -\log(1 - s(\beta_0 + \beta_1 x_i)) + \sum_{y_i=1} -\log(s(\beta_0 + \beta_1 x_i))$$

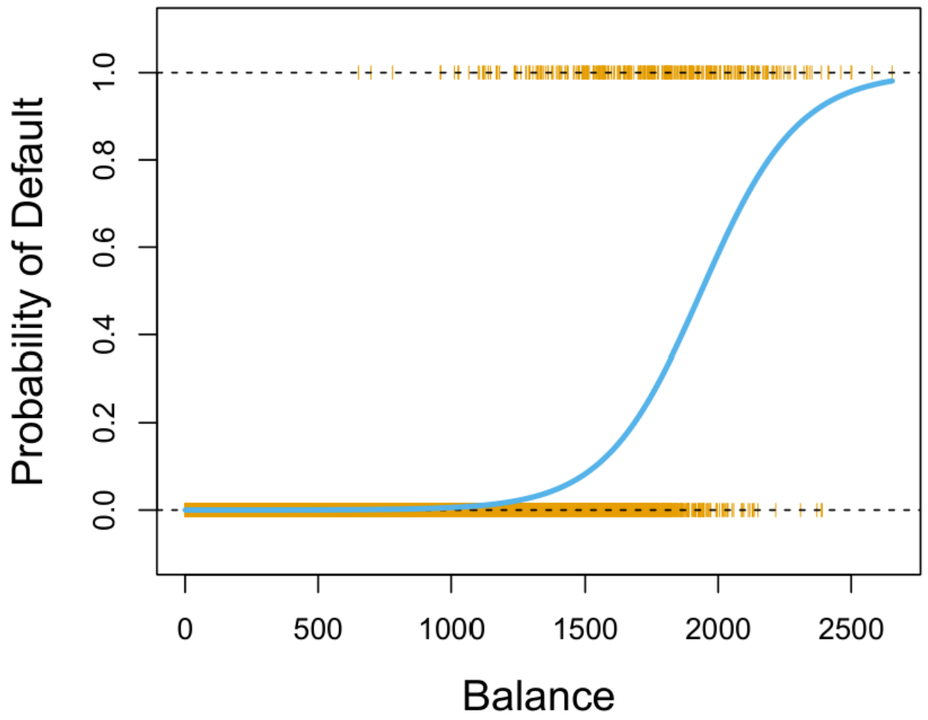


Logistic regression

Model: $\mathbb{P}(Y=1) = s(\beta_0 + \beta_1 X)$ with $s : t \mapsto \frac{e^t}{1+e^t}$ **(sigmoid)**

Minimize negative loglikelihood:

$$NL(\beta) = \sum_{y_i=0} -\log(1 - s(\beta_0 + \beta_1 x_i)) + \sum_{y_i=1} -\log(s(\beta_0 + \beta_1 x_i))$$



- When p predictors $X_1, \ldots, X_p$:

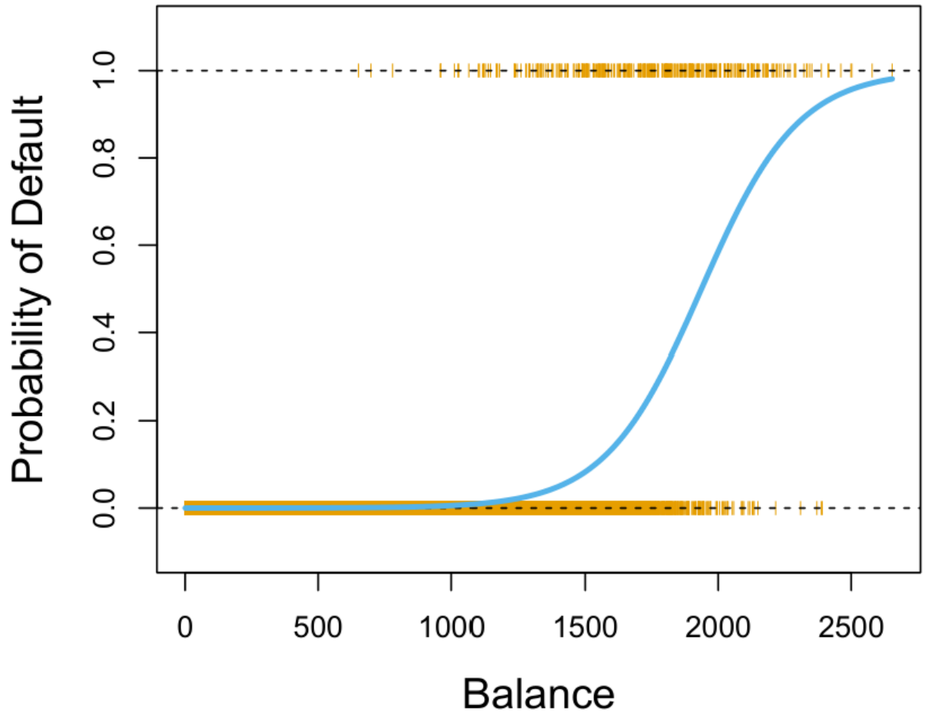
$$\beta_0 + \beta_1 X \longrightarrow \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Logistic regression

Model: $\mathbb{P}(Y=1) = s(\beta_0 + \beta_1 X)$ with $s : t \mapsto \frac{e^t}{1+e^t}$ **(sigmoid)**

Minimize negative loglikelihood:

$$NL(\beta) = \sum_{y_i=0} -\log(1 - s(\beta_0 + \beta_1 x_i)) + \sum_{y_i=1} -\log(s(\beta_0 + \beta_1 x_i))$$



- When p predictors $X_1, \ldots, X_p$:

$$\beta_0 + \beta_1 X \longrightarrow \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

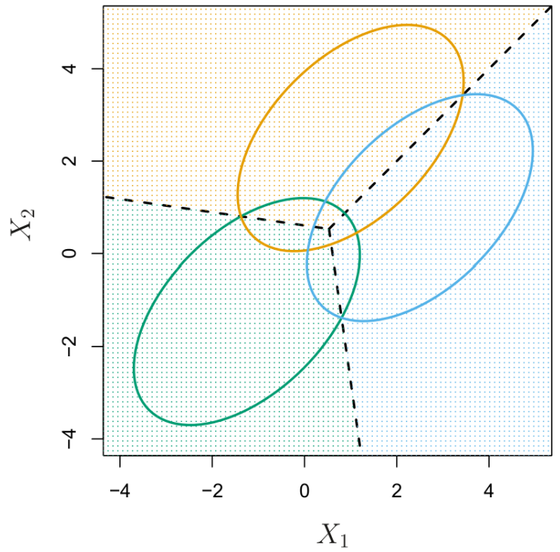
- When k classes with $k > 1$

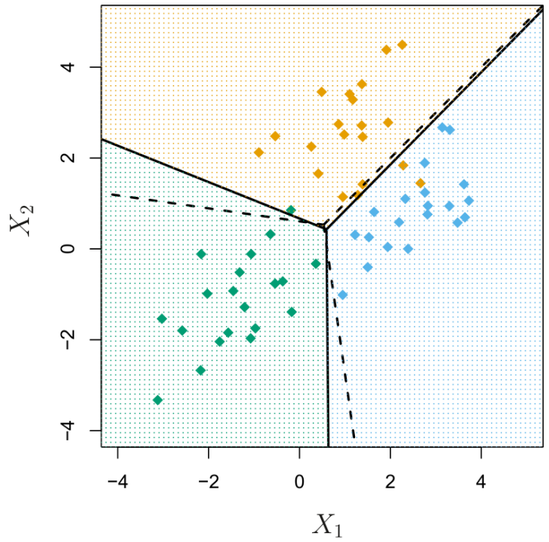
Look for the entries of a matrix

$\beta = (\beta_{h,i})_{h \in [k], i \in [p]} \in \mathbb{R}^{k \times p+1}$ with:

$$\text{Model: } \mathbb{P}(Y = l) = \frac{e^{\beta_{l,0} + \beta_{l,1} X_1 + \cdots + \beta_{l,p} X_p}}{\sum_{h=1}^{k} e^{\beta_{h,0} + \cdots + \beta_{h,p} X_p}}$$

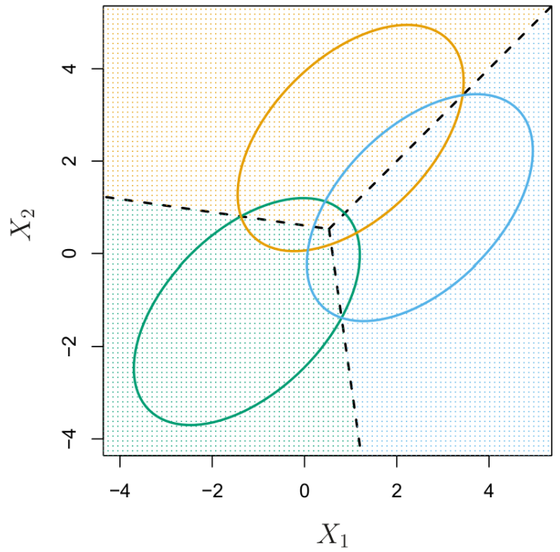
Linear discriminant analysis

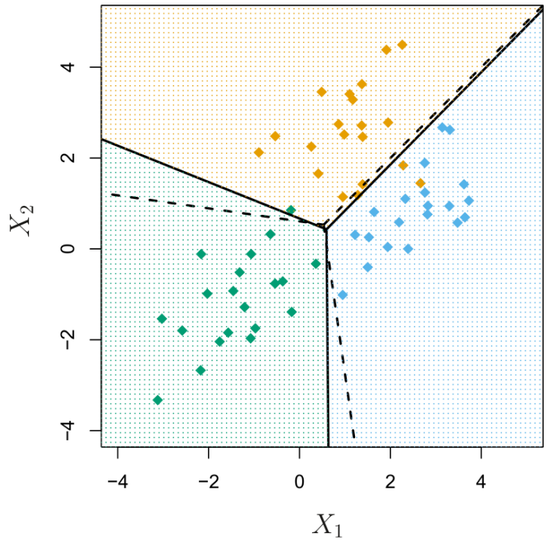




Bayes Rule: $\mathbb{P}(Y = l \mid X = x)\mathbb{P}(X = x) = \mathbb{P}(X = x \mid Y = l)\ \mathbb{P}(Y = l)$

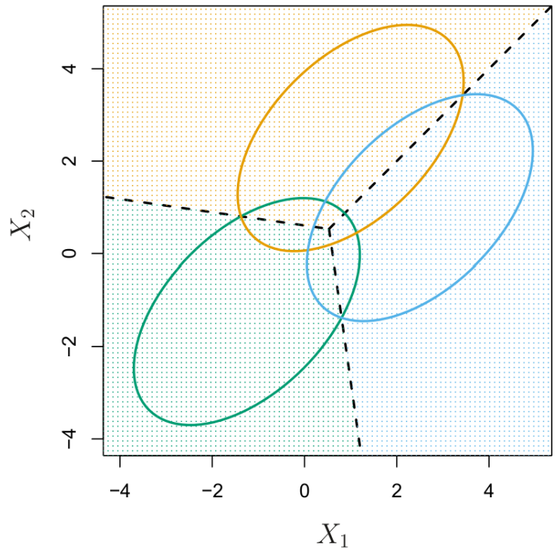
Linear discriminant analysis

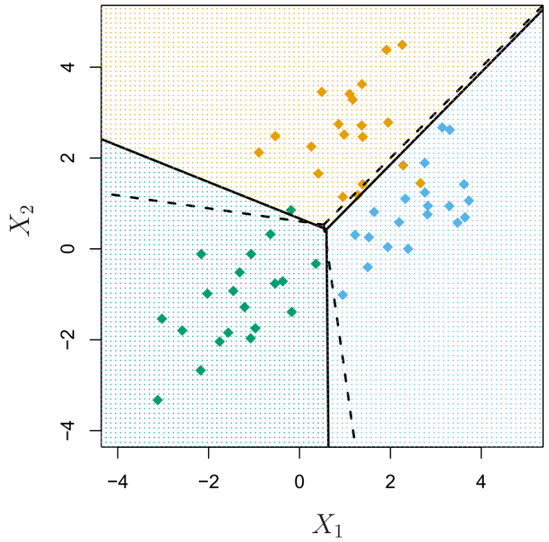




Bayes Rule: $\underbrace{\mathbb{P}(Y = l \mid X = x)}_{\text{our objective}} \mathbb{P}(X = x) = \underbrace{\mathbb{P}(X = x \mid Y = l)}_{\equiv f_l(x)} \underbrace{\mathbb{P}(Y = l)}_{\equiv \pi_l}$

Linear discriminant analysis

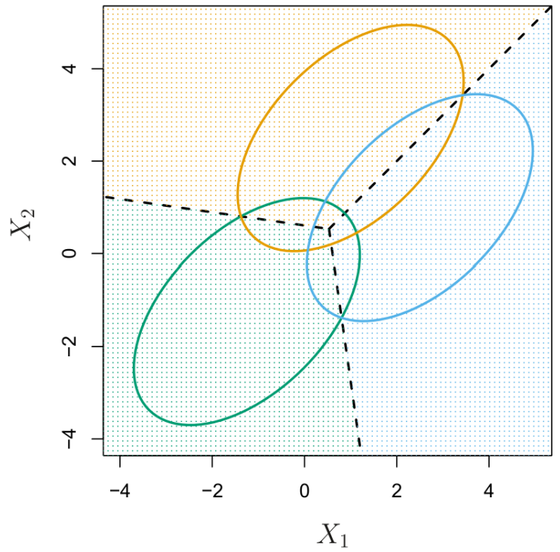


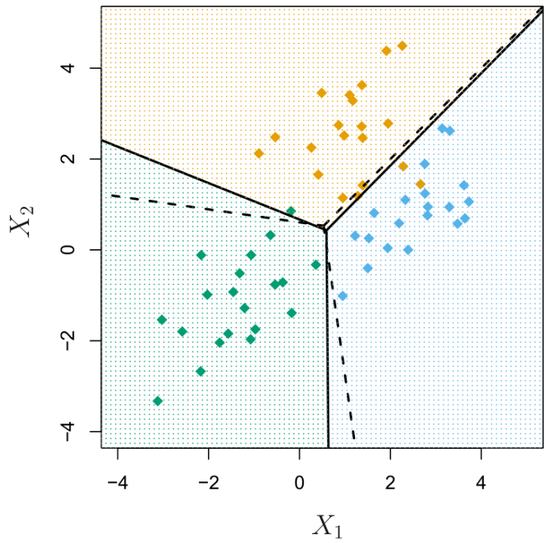


Bayes Rule: $\underbrace{\mathbb{P}(Y = l \mid X = x)}_{\text{our objective}} \mathbb{P}(X = x) = \underbrace{\mathbb{P}(X = x \mid Y = l)}_{\equiv f_l(x)} \underbrace{\mathbb{P}(Y = l)}_{\equiv \pi_l}$

$$\mathbb{P}(X = x) \sum_{l=1}^{k} \mathbb{P}(Y = l \mid X = x) = \sum_{l=1}^{k} f_l(x)\pi_l$$

Linear discriminant analysis



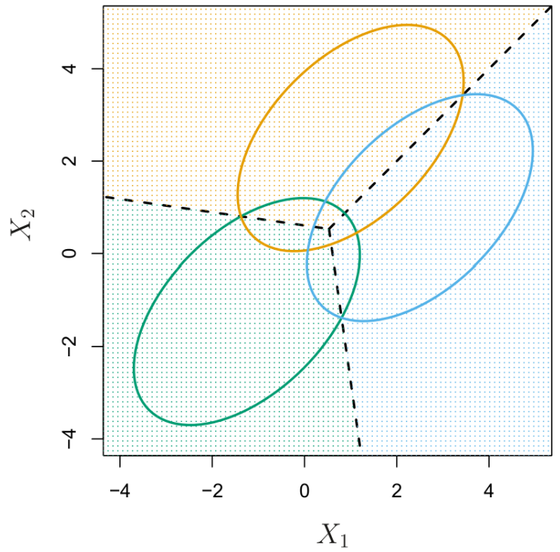


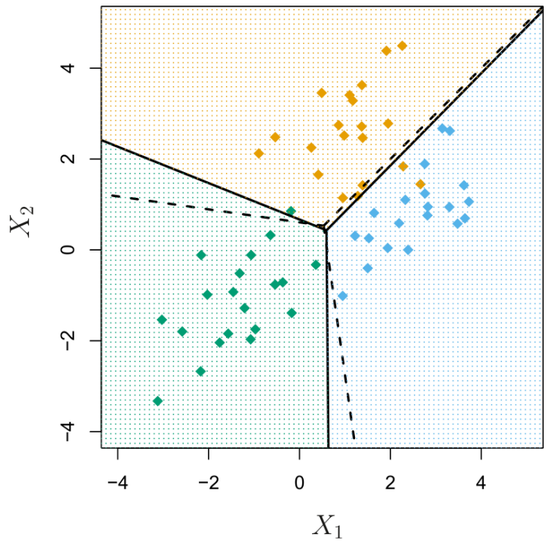
Bayes Rule: $\underbrace{\mathbb{P}(Y = l \mid X = x)}_{\text{our objective}} \mathbb{P}(X = x) = \underbrace{\mathbb{P}(X = x \mid Y = l)}_{\equiv f_l(x)} \underbrace{\mathbb{P}(Y = l)}_{\equiv \pi_l}$

$$\mathbb{P}(X = x) \underbrace{\sum_{l=1}^{k} \mathbb{P}(Y = l \mid X = x)}_{=1} = \sum_{l=1}^{k} f_l(x)\pi_l$$

Finally: $\mathbb{P}(Y = l \mid X = x) = \frac{f_l(x)\pi_l}{\sum_{h=1}^{k} f_h(x)\pi_h}$

Linear discriminant analysis





Bayes Rule: $\underbrace{\mathbb{P}(Y = l \mid X = x)}_{\text{our objective}} \mathbb{P}(X = x) = \underbrace{\mathbb{P}(X = x \mid Y = l)}_{\equiv f_l(x)} \underbrace{\mathbb{P}(Y = l)}_{\equiv \pi_l}$

$$\mathbb{P}(X = x) \underbrace{\sum_{l=1}^{k} \mathbb{P}(Y = l \mid X = x)}_{=1} = \sum_{l=1}^{k} f_l(x)\pi_l$$

Finally: $\mathbb{P}(Y = l \mid X = x) = \dfrac{f_l(x)\pi_l}{\sum_{h=1}^{k} f_h(x)\pi_h}$

In the case of known laws $X \mid Y = l \sim \mathcal{N}(\mu_l, \Sigma_l)$,

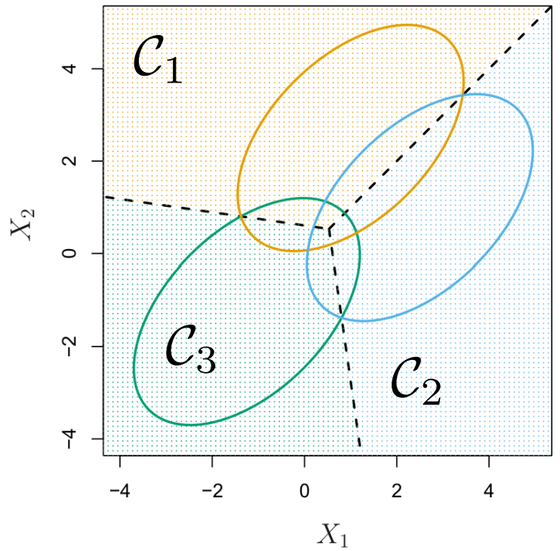
$$i.e.: \quad f_l(x) = \frac{\exp^{-\frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)}}{\sqrt{(2\pi)^p \det(\Sigma_l)}}$$

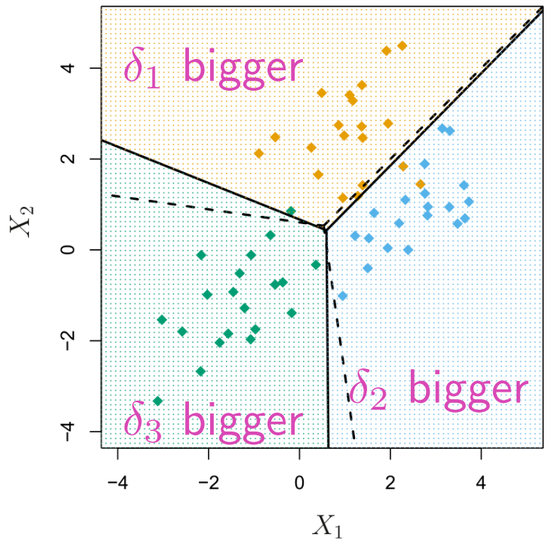
Given data x, look for $l \in [k]$ that maximizes $\mathbb{P}(Y = l \mid X = x)$

If $\Sigma_1 = \cdots = \Sigma_k$: $\quad l^{\text{th}}$ discriment: $\delta_l(x) = x^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l + \log(\pi_l)$

$\rightarrow$ Linear discriminant analysis

Linear discriminant analysis





Bayes Rule: $\underbrace{\mathbb{P}(Y = l \mid X = x)}_{\text{our objective}} \mathbb{P}(X = x) = \underbrace{\mathbb{P}(X = x \mid Y = l)}_{\equiv f_l(x)} \underbrace{\mathbb{P}(Y = l)}_{\equiv \pi_l}$

$$\mathbb{P}(X = x) \underbrace{\sum_{l=1}^k \mathbb{P}(Y = l \mid X = x)}_{=1} = \sum_{l=1}^k f_l(x)\pi_l$$

Finally: $\mathbb{P}(Y = l \mid X = x) = \dfrac{f_l(x)\pi_l}{\sum_{h=1}^k f_h(x)\pi_h}$

In the case of known laws $X \mid Y = l \sim \mathcal{N}(\mu_l, \Sigma_l)$,

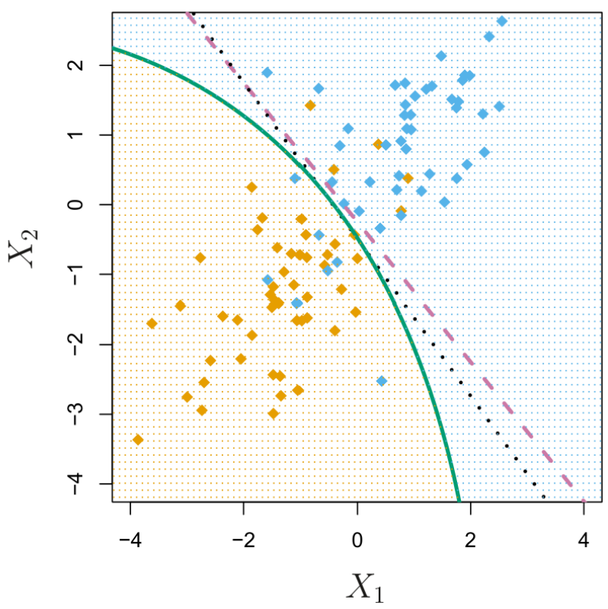
$$i.e.: \quad f_l(x) = \frac{\exp^{-\frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)}}{\sqrt{(2\pi)^p \det(\Sigma_l)}}$$

Given data x, look for $l \in [k]$ that maximizes $\mathbb{P}(Y = l \mid X = x)$

If $\Sigma_1 = \cdots = \Sigma_k$: $\quad l^{\text{th}}$ discriment: $\delta_l(x) = x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + \log(\pi_l)$

$\rightarrow$ Linear discriminant analysis

Quadratic discriminant analysis



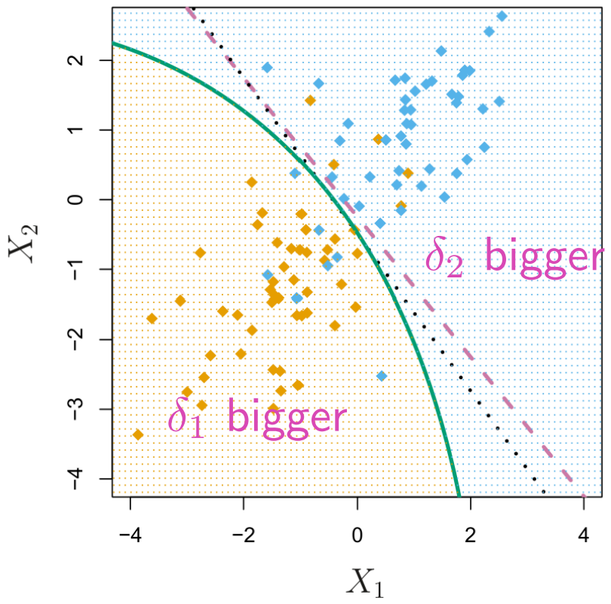
Model: $\mathbb{P}(Y = l \mid X = x) = \frac{f_j(x)\pi_l}{\sum_{h=1}^{k} f_h(x)\pi_h}$

with : $f_l(x) = \frac{\exp^{-\frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)}}{\sqrt{(2\pi)^p \det(\Sigma_l)}}$

- **Linear discriminent analysis:** $(\Sigma_1 = \cdots = \Sigma_k \equiv \Sigma)$

$$\delta_l(x) = x^T\Sigma^{-1}\mu_l - \tfrac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + \log(\pi_l)$$

Quadratic discriminant analysis



Model: $\mathbb{P}(Y = l \mid X = x) = \frac{f_j(x)\pi_l}{\sum_{h=1}^{k} f_h(x)\pi_h}$

with : $f_l(x) = \frac{\exp^{-\frac{1}{2}(x-\mu_l)^T\Sigma_l^{-1}(x-\mu_l)}}{\sqrt{(2\pi)^p \det(\Sigma_l)}}$

- **Linear discriminent analysis:** $(\Sigma_1 = \cdots = \Sigma_k \equiv \Sigma)$

$$\delta_l(x) = x^T\Sigma^{-1}\mu_l - \tfrac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + \log(\pi_l)$$

- **Quadratic discriminent analysis:** (the covariances are different)

$$\delta_l(x) = -\tfrac{1}{2}x^T\Sigma_l x + x^T\Sigma_l^{-1}\mu_l - \tfrac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + \log(\pi_l) - \tfrac{1}{2}\log(\det(\Sigma_l))$$

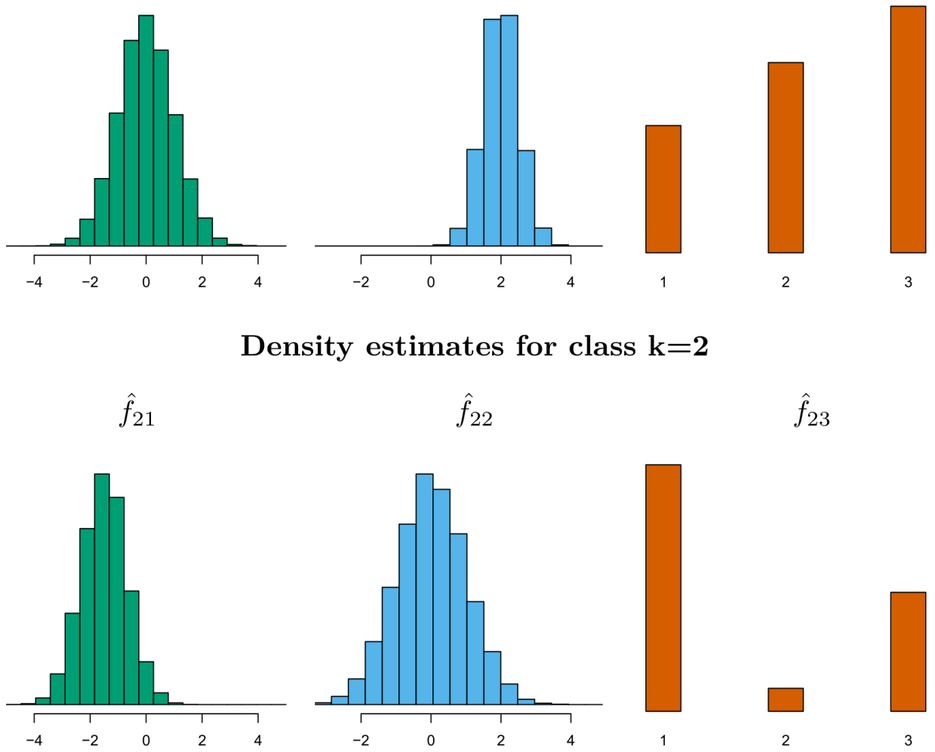
$\longrightarrow$ More flexible ! **but** More precise ?

In practice, $(\mu_l)_{l\in[k]}$, $(\Sigma_l)_{l\in[k]}$ estimated with $\forall l \in [k]$:

$$\hat{\mu}_l \equiv \frac{\pi_l}{n}\sum_{i=1,y_i=l}^{n} x_i \quad \text{and} \quad \hat{\Sigma}_l \equiv \frac{\pi_l}{n}\sum_{i=1,y_i=l}^{n}(x_i-\mu_l)(x_i-\mu_l)^T$$

Pb: Estimation of covariance highly sensitive to noise.

Naive Bayes



Always works but lacks some flexibility

Strong hypothesis: All predictors $X_1, \dots, X_p$ idpts.

Implies all Σ_l diagonal and $f_l(x) = f_l(x_1) \cdots f_l(x_p)$

(Increase bias but decreases variance)

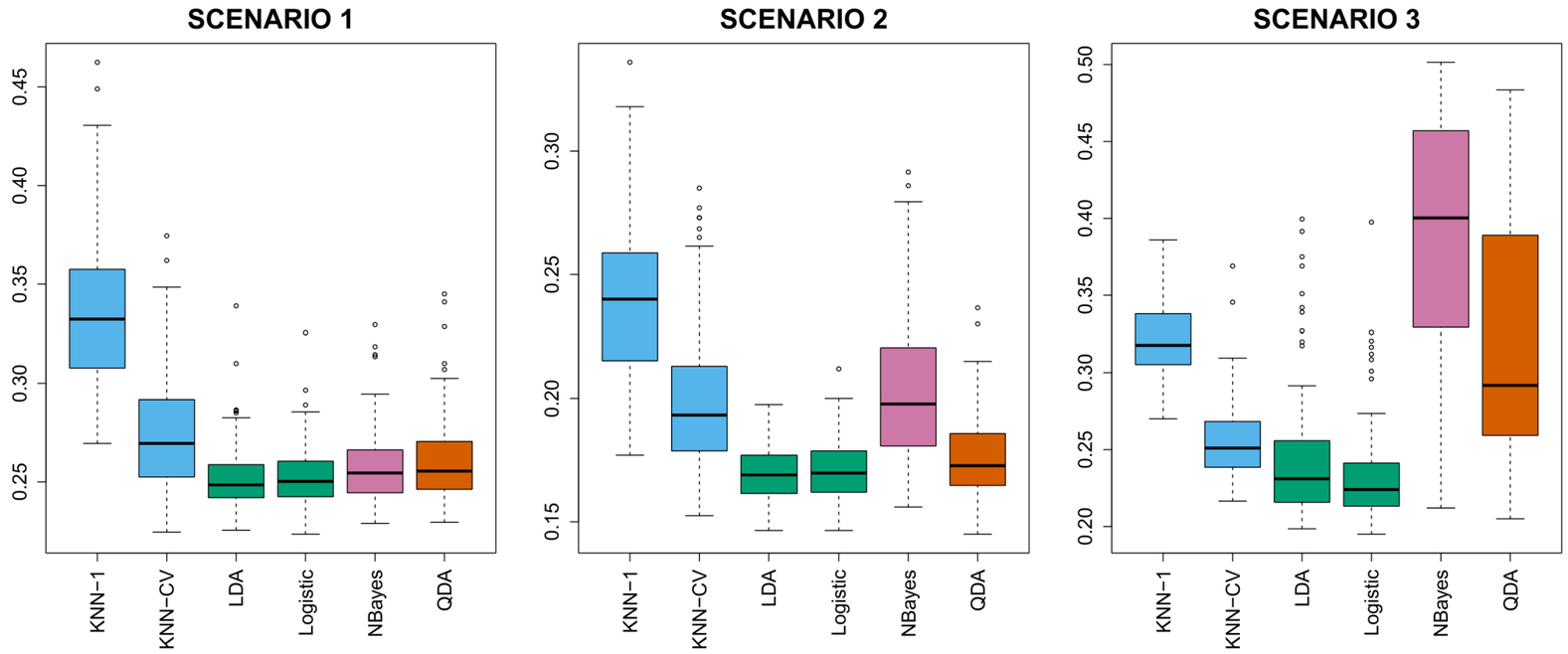
Estimation of the $f_{l,i}$, $l \in [k]$, $i \in [p]$:

- Gaussian assumption $f_{l,i}(x) = \frac{e^{-\frac{(x-\mu_{l,i})^2}{2\sigma_{l,i}^2}}}{\sqrt{2\pi}\sigma_{l,i}}$
- Directly take histogram $f_{l,i}(x) = \sum_{b \in \text{bins}} 1_{x \in b} p_b$

Discriminant analysis again, given $x \in \mathbb{R}^p$ choose l that maximizes:

$$\mathbb{P}(Y = l \mid X = x) = \frac{f_l(x)\pi_l}{\sum_{h=1}^{k} f_h(x)\pi_h}$$

Comparison

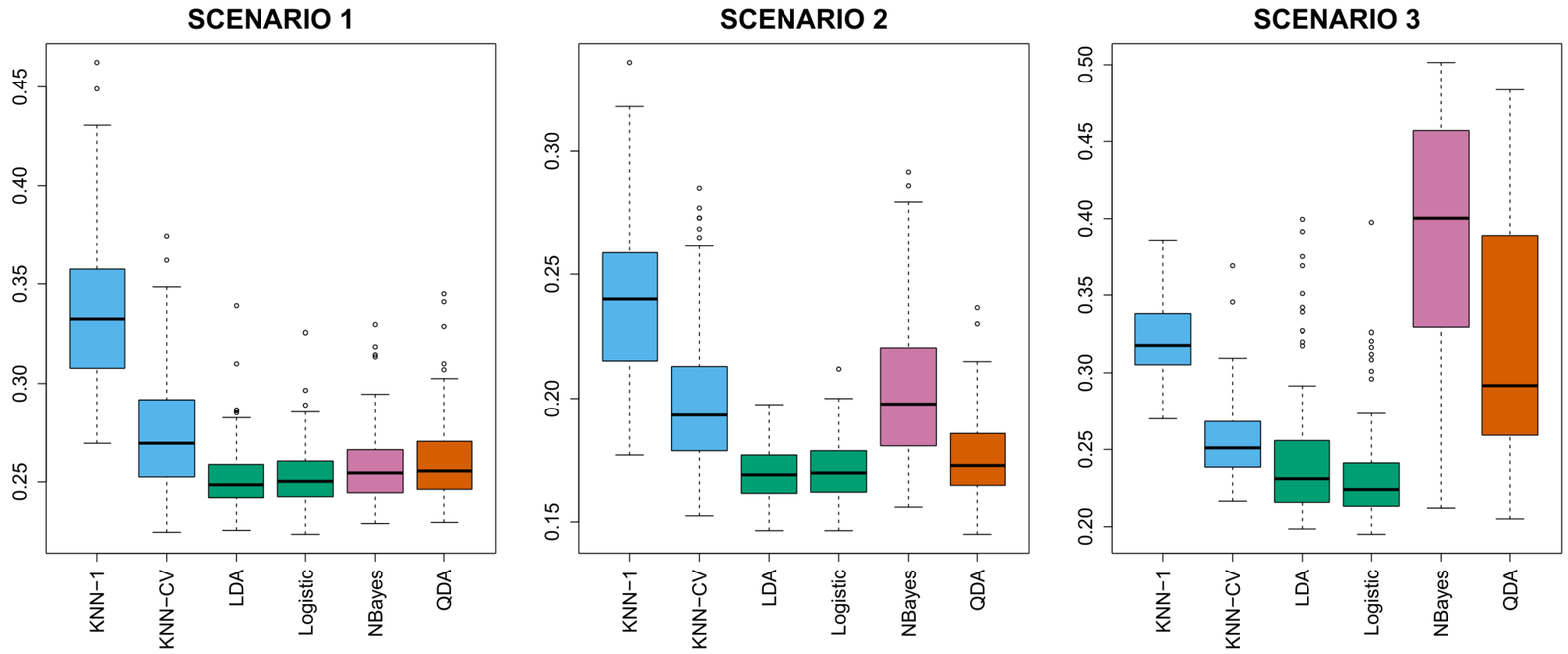


Senario 1:

- 20 training Gaussian observation
- Observations uncorrelated

$\rightarrow$ k-means too deterministic, QDA too flexible

Comparison



Senario 1:

- 20 training Gaussian observation
- Observations uncorrelated

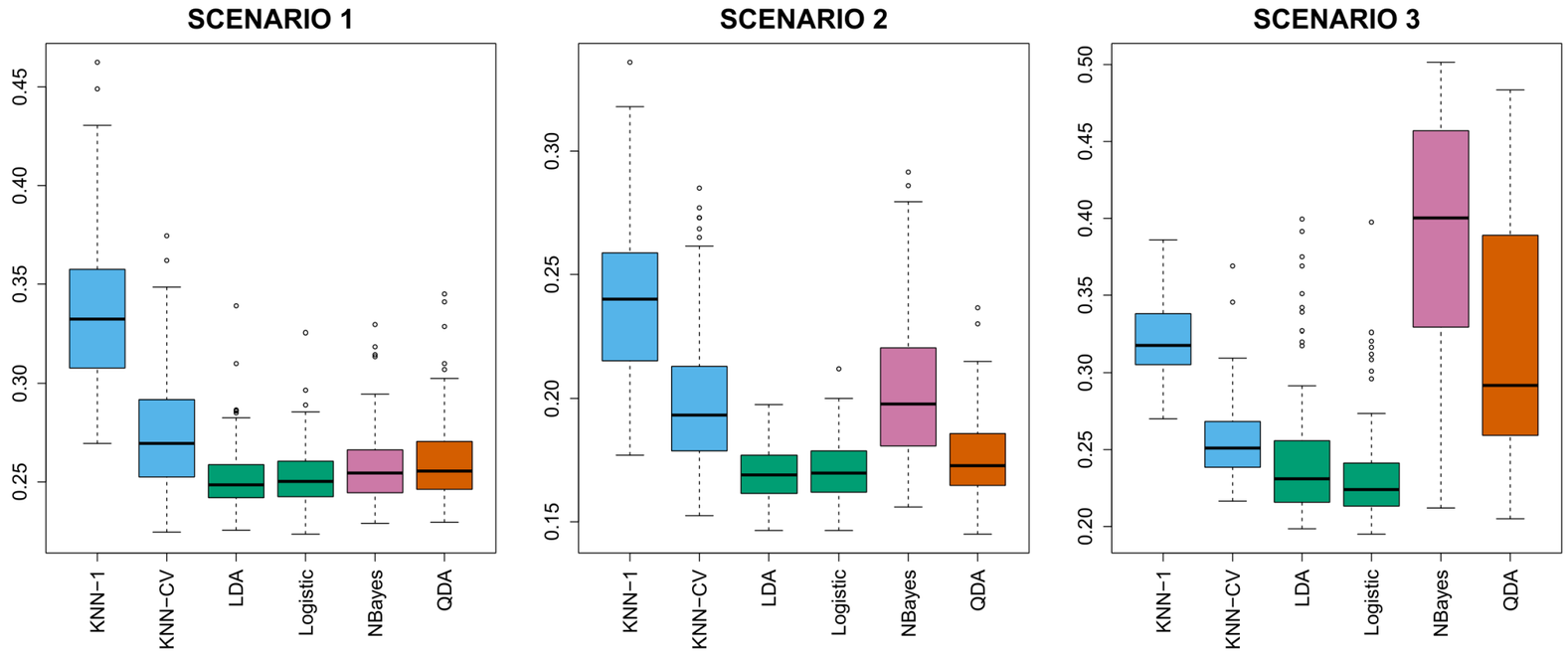
$\rightarrow$ k-means too deterministic, QDA too flexible

Senario 2:

- 20 training Gaussian observation
- Correlation of -0.5 between predictors

$\rightarrow$ Naive bayes independence assumption not satisfied

Comparison



Senario 3:

- 20 training t-distributed observation
- Correlation of -0.5 between predictors

$\rightarrow$ LDA, QDA assumtions violated

Senario 1:

- 20 training Gaussian observation
- Observations uncorrelated

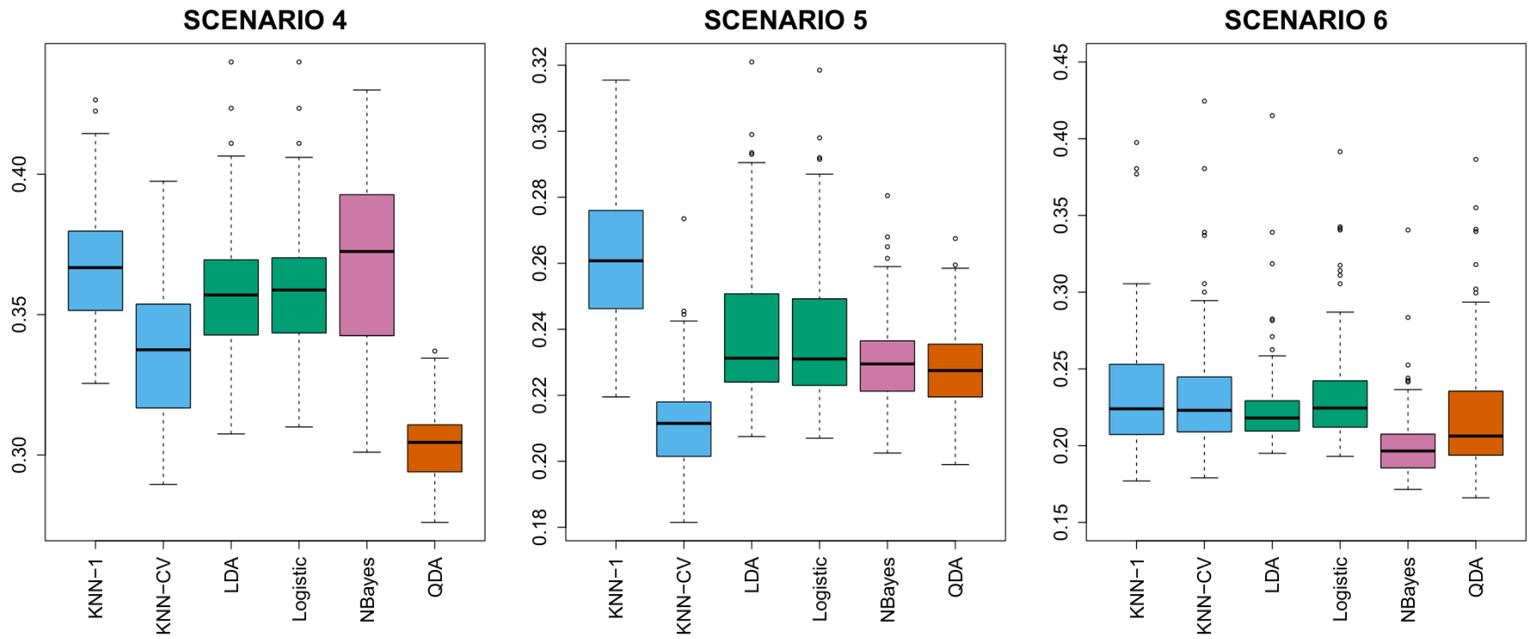
$\rightarrow$ k-means too deterministic, QDA too flexible

Senario 2:

- 20 training Gaussian observation
- Correlation of -0.5 between predictors

$\rightarrow$ Naive bayes independence assumption not satisfied

Comparison

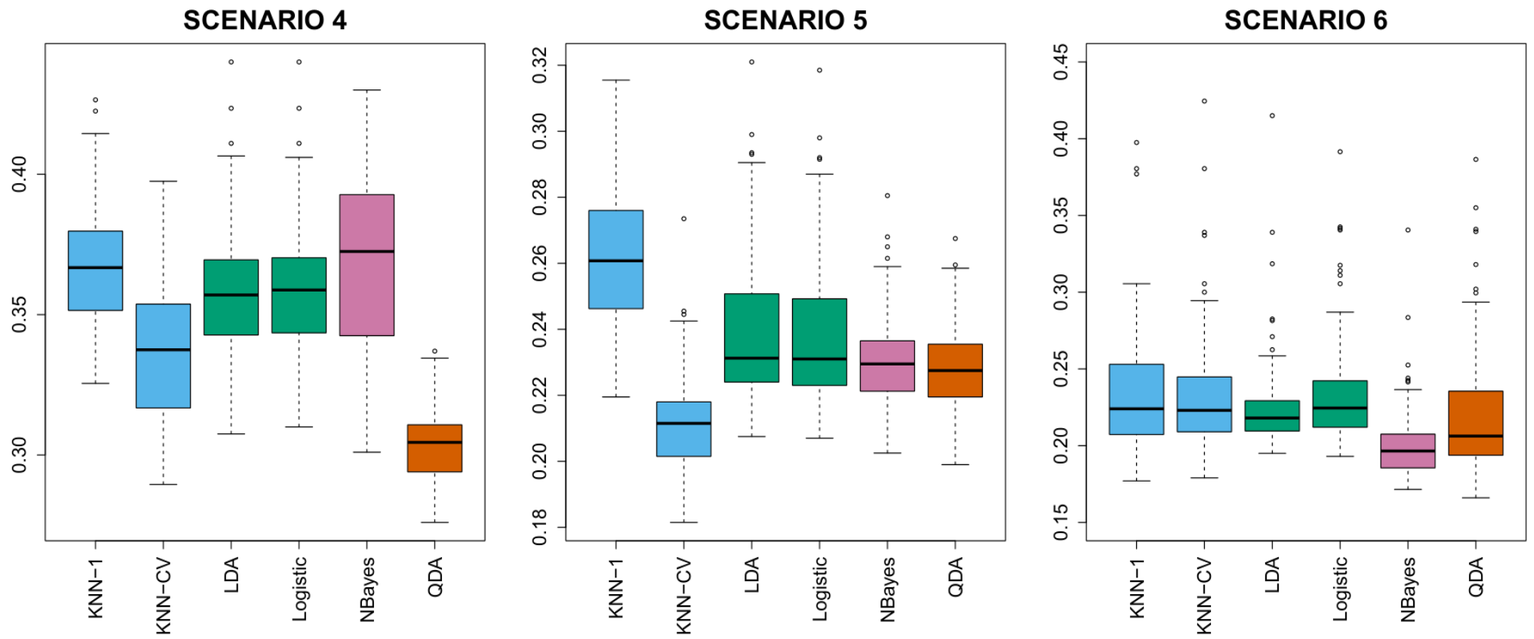


Senario 4:

- 20 training Gaussian observation
- Correlations between predictors different between two classes

$\rightarrow$ perfect fit of QDA method (decision boundary non linear)

Comparison



Senario 4:

- 20 training Gaussian observation
- Correlations between predictors different between two classes

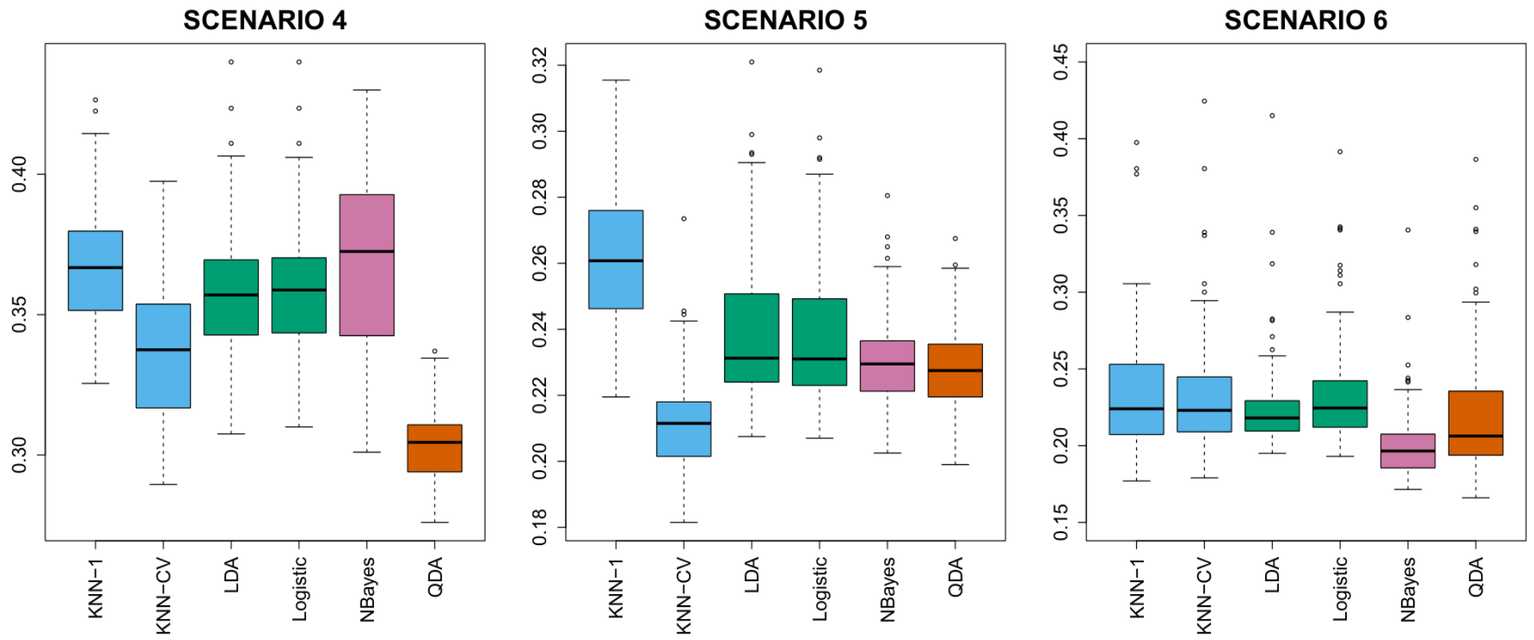
$\rightarrow$ perfect fit of QDA method (decision boundary non linear)

Senario 5:

- 20 training Gaussian observation
- Response Y logistic image of non linear transformation of predictors X

$\rightarrow$ k-means robust to this difficult setting

Comparison



Senario 3:

- 6 training Gaussian observations
- no correlation between predictors but variance different between classes

$\rightarrow$ independence is a perfect fit for naive Bayes

$\rightarrow$ variance too high for QDA

Senario 4:

- 20 training Gaussian observation
- Correlations between predictors different between two classes

$\rightarrow$ perfect fit of QDA method (decision boundary non linear)

Senario 5:

- 20 training Gaussian observation
- Response Y logistic image of non linear transformation of predictors X

$\rightarrow$ k-means robust to this difficult setting

More general methods: Empirical risk minimization

Here, no specific model for the data (flexible parametric method).

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} l(h_w(x_i), y_i) + \qquad \lambda r(y_i), \qquad w \in \mathcal{R}^q$$

More general methods: Empirical risk minimization

Here, no specific model for the data (flexible parametric method).

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} \underbrace{l(h_w(x_i), y_i)}_{Loss} + \underbrace{\lambda r(y_i)}_{Regularization} \qquad w \in \mathcal{R}^q$$

More general methods: Empirical risk minimization

Here, no specific model for the data (flexible parametric method).

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} \underbrace{l(h_w(x_i), y_i)}_{Loss} + \underbrace{\lambda r(y_i)}_{Regularization} \qquad w \in \mathcal{R}^q$$

h_w: hypothesis
w: parameter

More general methods: Empirical risk minimization

Here, no specific model for the data (flexible parametric method).

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} \underbrace{l(h_w(x_i), y_i)}_{Loss} + \underbrace{\lambda r(y_i)}_{Regularization} \qquad w \in \mathcal{R}^q$$

h_w: hypothesis
w: parameter

r: regularizing loss, will be studied later in the lecture on interpolation vs extrapolation.

More general methods: Empirical risk minimization

Here, no specific model for the data (flexible parametric method).

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} \underbrace{l(h_w(x_i), y_i)}_{Loss} + \underbrace{\lambda r(y_i)}_{Regularization} \qquad w \in \mathcal{R}^q$$

h_w: hypothesis
w: parameter

r: regularizing loss, will be studied later in the lecture on interpolation vs extrapolation.

Example

Ridge regression: Minimize $\frac{1}{n} \sum_{i=1}^{n} \|\beta x_i - y_i\|^2 + \lambda \|\beta\|^2, \quad \beta \in \mathbb{R}^p$

More general methods: Empirical risk minimization

Here, no specific model for the data (flexible parametric method).

$$\text{Minimize} \frac{1}{n}\sum_{i=1}^{n} \underbrace{l(h_w(x_i), y_i)}_{Loss} + \underbrace{\lambda r(y_i)}_{Regularization} \qquad w \in \mathcal{R}^q$$

h_w: hypothesis
w: parameter

r: regularizing loss, will be studied later in the lecture on interpolation vs extrapolation.

Example

Ridge regression: Minimize $\frac{1}{n}\sum_{i=1}^{n} \|\beta x_i - y_i\|^2 + \lambda\|\beta\|^2, \quad \beta \in \mathbb{R}^p$

Lasso: Minimize $\frac{1}{n}\sum_{i=1}^{n} \|\beta x_i - y_i\|^2 + \lambda\|\beta\|_1, \quad \beta \in \mathbb{R}^p$

More general methods: Empirical risk minimization

Here, no specific model for the data (flexible parametric method).

Minimize $\frac{1}{n}\sum_{i=1}^{n}\underbrace{l(h_w(x_i),y_i)}_{Loss}+\underbrace{\lambda r(y_i)}_{Regularization}$ $\quad w\in\mathcal{R}^q$

h_w: hypothesis
w: parameter

r: regularizing loss, will be studied later in the lecture on interpolation vs extrapolation.

Example

Ridge regression: Minimize $\frac{1}{n}\sum_{i=1}^{n}\|\beta x_i-y_i\|^2+\lambda\|\beta\|^2, \quad \beta\in\mathbb{R}^p$

Lasso: Minimize $\frac{1}{n}\sum_{i=1}^{n}\|\beta x_i-y_i\|^2+\lambda\|\beta\|_1, \quad \beta\in\mathbb{R}^p$

Support vector machines **(SVM)**: Minimize $\frac{1}{n}\sum_{i=1}^{n}\max(1-(\beta^T x_i+b)y_i)+\lambda\|\beta\|^2, \quad \beta\in\mathbb{R}^p.$

More general methods: Empirical risk minimization

Here, no specific model for the data (flexible parametric method).

Minimize $\frac{1}{n}\sum_{i=1}^{n}\underbrace{l(h_w(x_i),y_i)}_{Loss}+\underbrace{\lambda r(y_i)}_{Regularization}$ $\quad w\in\mathcal{R}^q$

h_w: hypothesis
w: parameter

r: regularizing loss, will be studied later in the lecture on interpolation vs extrapolation.

Example

Ridge regression: Minimize $\frac{1}{n}\sum_{i=1}^{n}\|\beta x_i-y_i\|^2+\lambda\|\beta\|^2, \quad \beta\in\mathbb{R}^p$

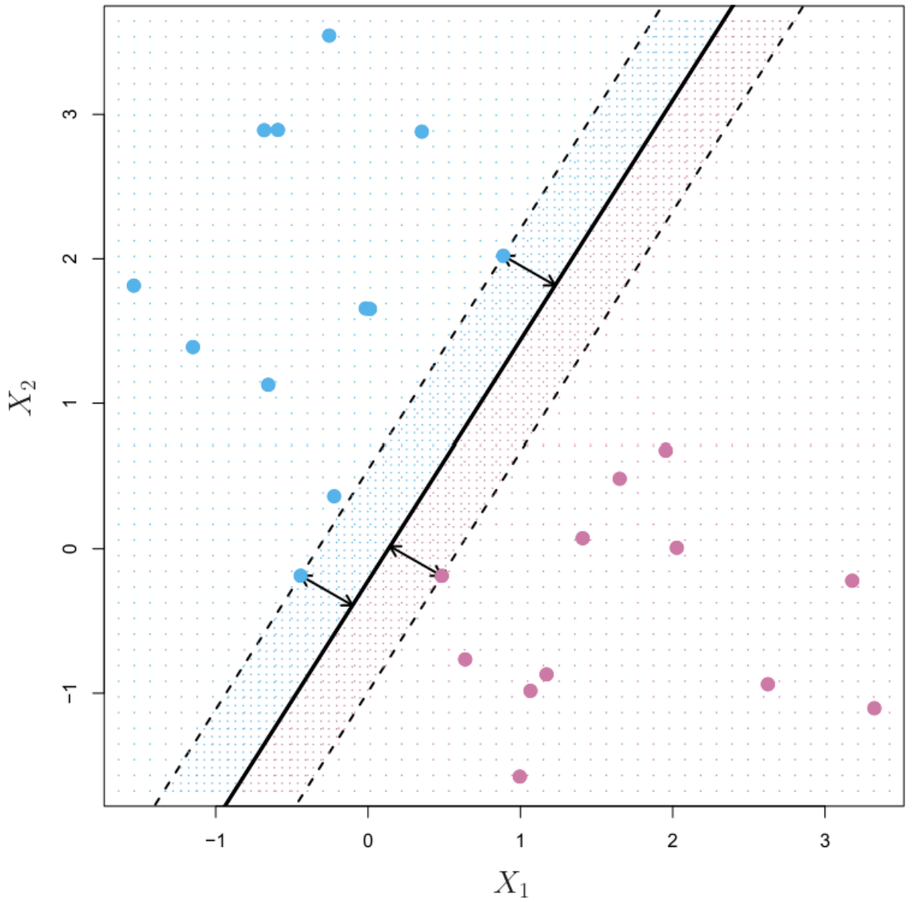
Lasso: Minimize $\frac{1}{n}\sum_{i=1}^{n}\|\beta x_i-y_i\|^2+\lambda\|\beta\|_1, \quad \beta\in\mathbb{R}^p$

Support vector machines **(SVM)**: Minimize $\frac{1}{n}\sum_{i=1}^{n}\max(1-(\beta^T x_i+b)y_i)+\lambda\|\beta\|^2, \quad \beta\in\mathbb{R}^p.$

Playing on losses can allow to go from classification to regression

Some words about "Support vector machines"

Support vector machines **(SVM)**: Minimize $\frac{1}{n}\sum_{i=1}^{n}\max(1-(\beta^T x_i+b)y_i)+\lambda\|\beta\|^2, \quad \beta\in\mathbb{R}^p.$

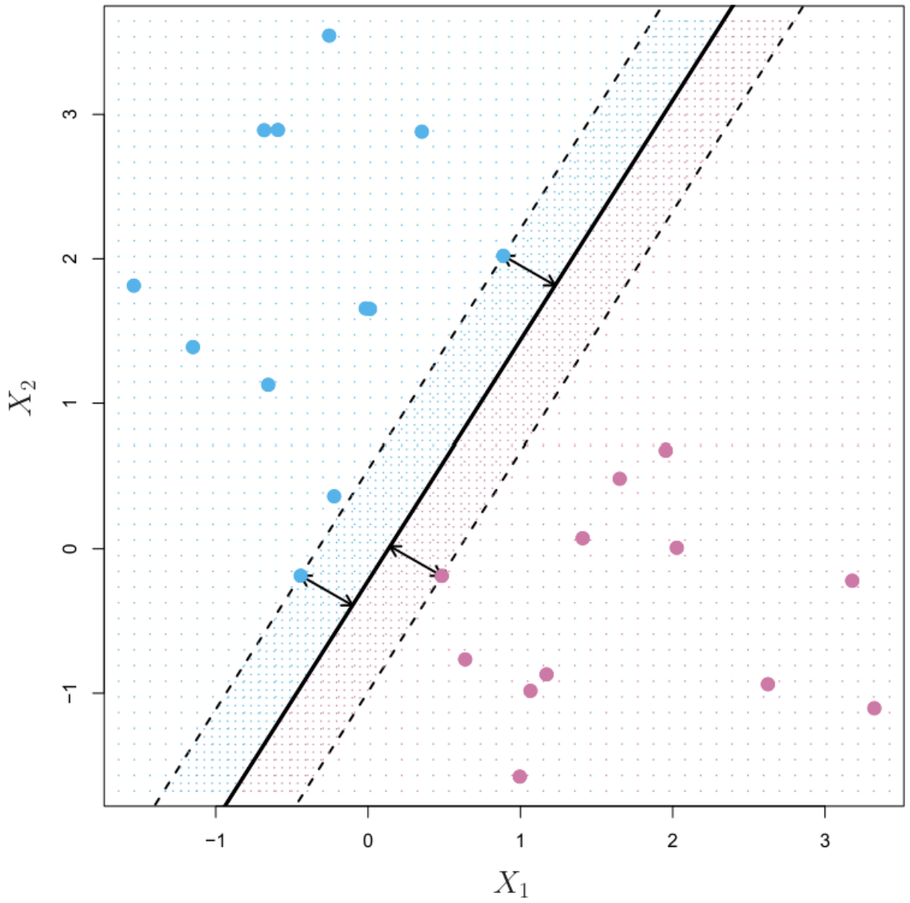


Closely related to so-called "Maximal Margin classifier", very popular from the 90's

Decision boundary linear $\rightarrow$ improvement with Kernel SVM:

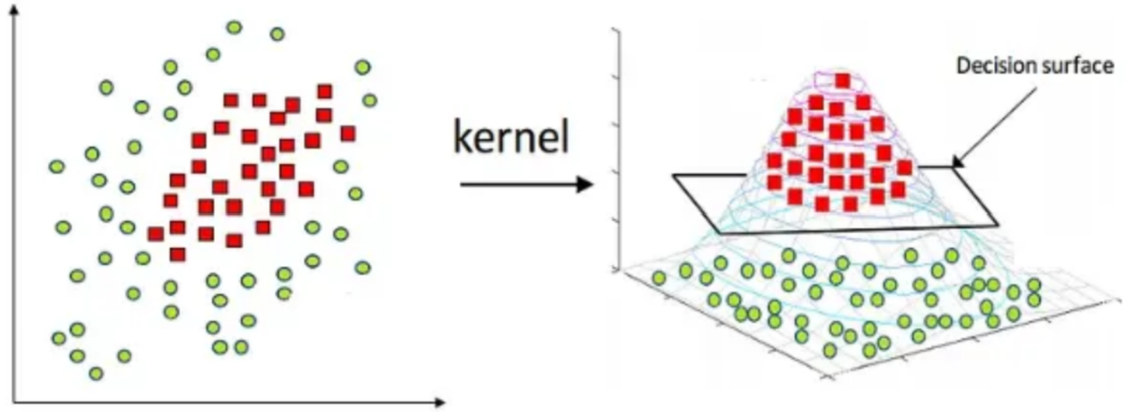
Some words about "Support vector machines"

Support vector machines **(SVM)**: Minimize $\frac{1}{n}\sum_{i=1}^{n}\max(1-(\beta^T x_i+b)y_i)+\lambda\|\beta\|^2, \quad \beta\in\mathbb{R}^p.$



Closely related to so-called "Maximal Margin classifier", very popular from the 90's

Decision boundary linear $\rightarrow$ improvement with Kernel SVM:

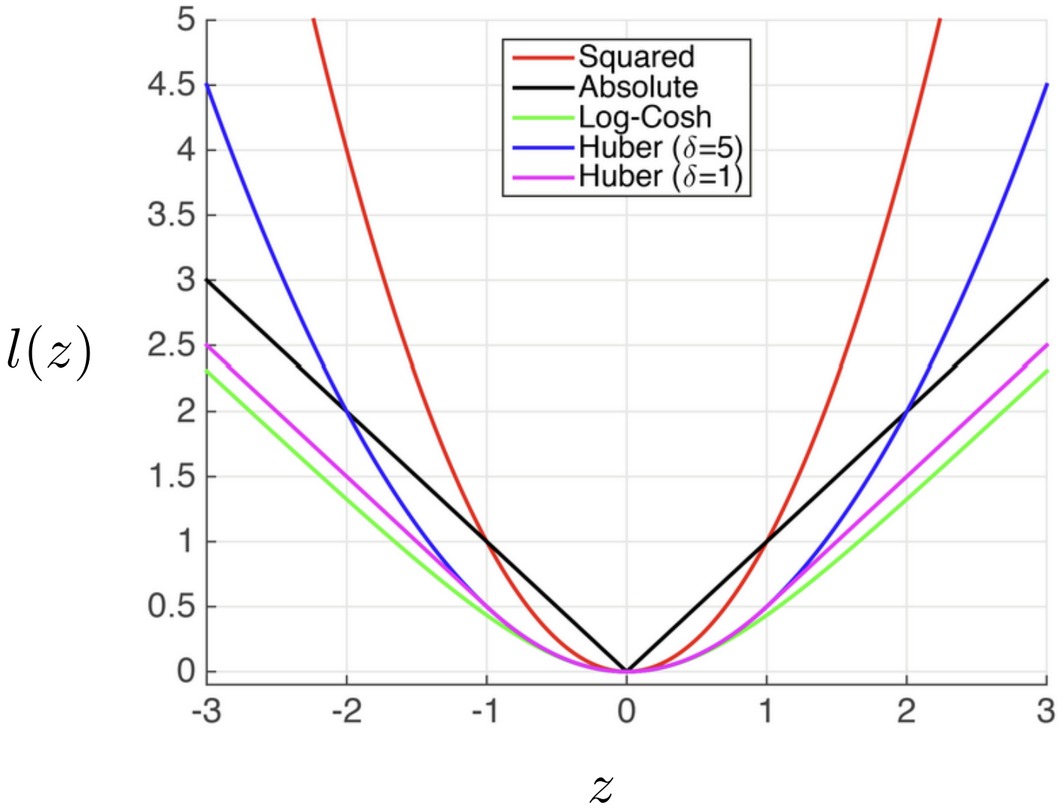


Regression losses $Y \in \mathbb{R}$

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} l(h_w(x_i) - y_i) + \lambda r(y_i) \qquad w \in \mathcal{R}^q \qquad h_w(x_i) - y_i \mapsto z$$

Regression losses $Y \in \mathbb{R}$

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} l(h_w(x_i) - y_i) + \lambda r(y_i) \qquad w \in \mathcal{R}^q \qquad h_w(x_i) - y_i \mapsto z$$

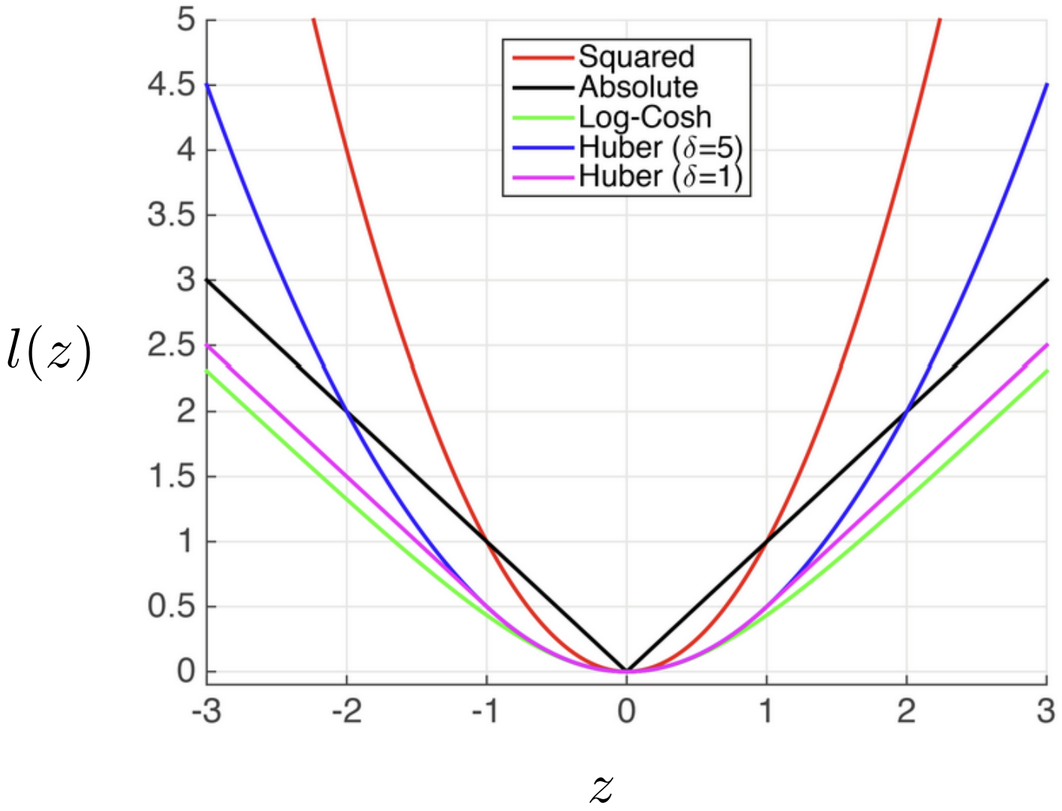


Squared loss: $l(z) = z^2$

Used for LSR, sensitive to outliers

Regression losses $Y \in \mathbb{R}$

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} l(h_w(x_i) - y_i) + \lambda r(y_i) \qquad w \in \mathcal{R}^q \qquad h_w(x_i) - y_i \mapsto z$$

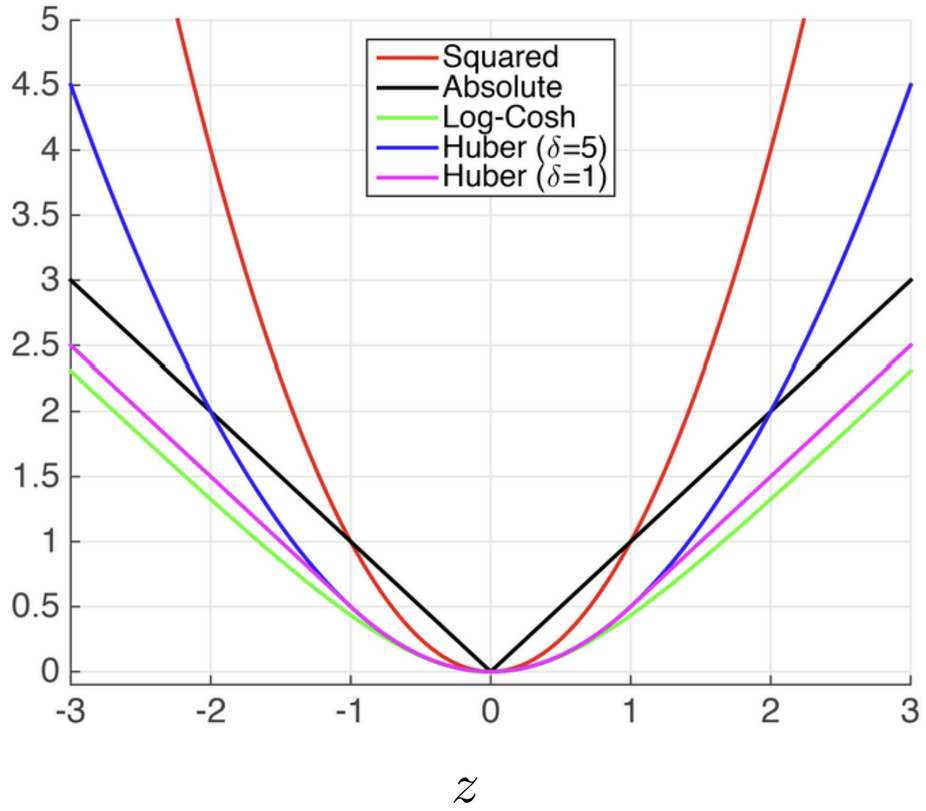


Squared loss: $l(z) = z^2$
Used for LSR, sensitive to outliers

Absolute loss: $l(z) = |z|$
Provides median labels, less sensitive to outliers, non differentiable in zero

Regression losses $Y \in \mathbb{R}$

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} l(h_w(x_i) - y_i) + \lambda r(y_i) \qquad w \in \mathcal{R}^q \qquad h_w(x_i) - y_i \mapsto z$$



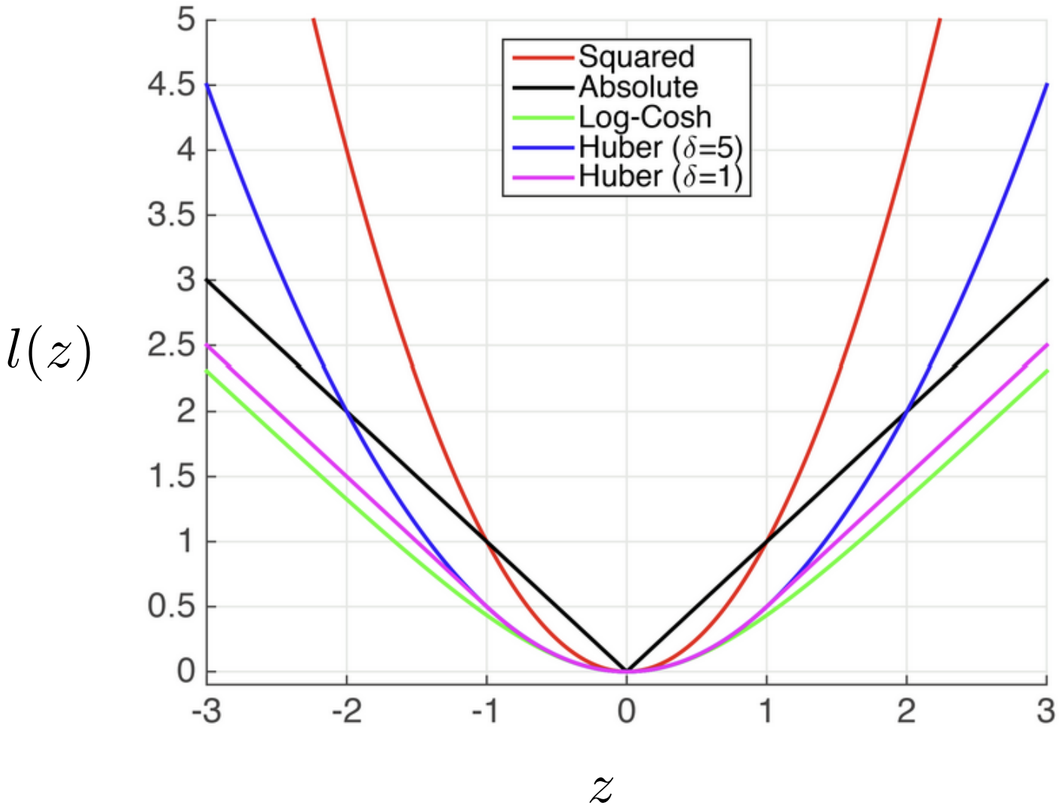
Squared loss: $l(z) = z^2$
Used for LSR, sensitive to outliers

Absolute loss: $l(z) = |z|$
Provides median labels, less sensitive to outliers, non differentiable in zero

Huber loss: $l(z) = z^2$ if $|z| < \delta$, $l(z) = 2\delta|z| - \delta^2$ othws.
Advantages of squared and absolute loss.

Regression losses $Y \in \mathbb{R}$

Minimize $\frac{1}{n}\sum_{i=1}^{n} l(h_w(x_i) - y_i) + \lambda r(y_i)$ $\quad w \in \mathcal{R}^q$ $\quad h_w(x_i) - y_i \mapsto z$



Squared loss: $l(z) = z^2$
Used for LSR, sensitive to outliers

Absolute loss: $l(z) = |z|$
Provides median labels, less sensitive to outliers, non differentiable in zero

Huber loss: $l(z) = z^2$ if $|z| < \delta$, $l(z) = 2\delta|z| - \delta^2$ othws.
Advantages of squared and absolute loss.

Log-cosh loss: $l(z) = \log(\cosh(z))$ (where $\cosh(z) \equiv \frac{e^z + e^{-z}}{2}$)
Like Huber loss but twice differentiable in 0.

Classification losses $Y \in \{-1, 1\}$

Minimize $\frac{1}{n}\sum_{i=1}^{n} l(h_w(x_i) \cdot y_i) + \lambda r(y_i)$ $\quad w \in \mathcal{R}^q$ $\quad h_w(x_i)y_i \mapsto z$

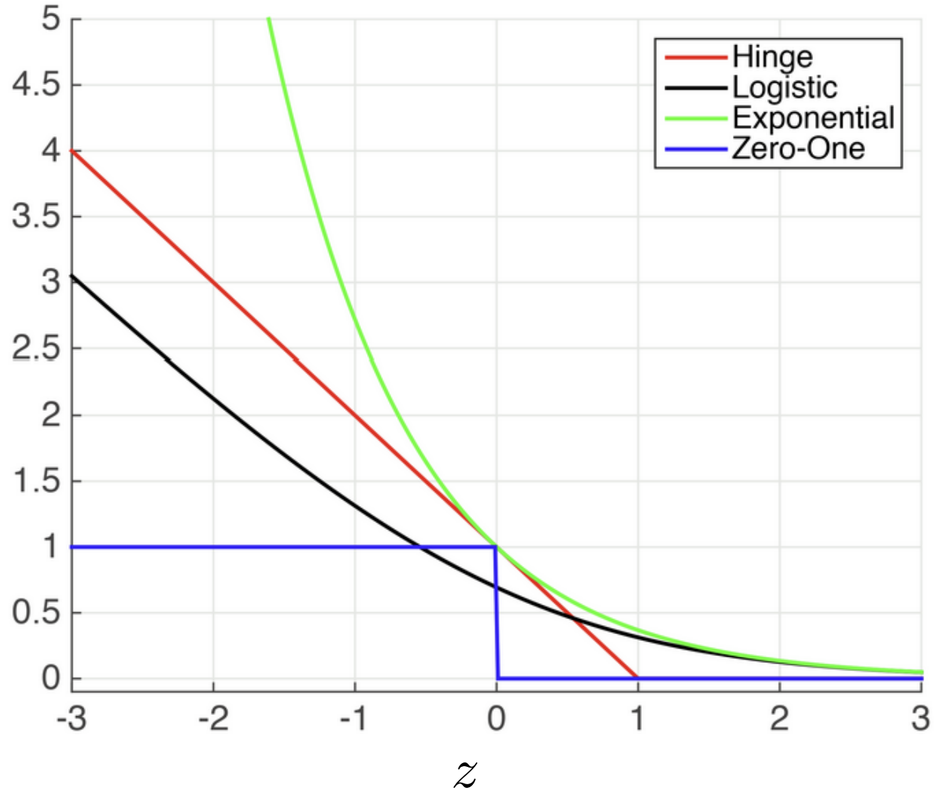
In logistic regression look for $\beta \in \mathbb{R}^p$ that minimizes:

$$-\sum_{y_i=1} \log\left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right) - \sum_{y_i=-1} \log\left(1 - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)$$

- $\log\left(1 - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right) = -l(\beta_i^x) = -l(\beta^T x_i \cdot y_i)$ if $y_i = 1$
- $\log\left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right) = \log\left(\frac{1}{e^{-\beta^T x_i}+1}\right) = -l(\beta^T x_i \cdot y_i)$ if $y_i = -1$.

Classification losses $Y \in \{-1, 1\}$

Minimize $\frac{1}{n}\sum_{i=1}^{n} l(h_w(x_i)\cdot y_i) + \lambda r(y_i)$ $\quad w \in \mathcal{R}^q$ $\quad h_w(x_i)y_i \mapsto z$



Logistic loss: $l(z) = \log(1 + e^{-z})$

Used for logistic regression, probabilistic interpretation

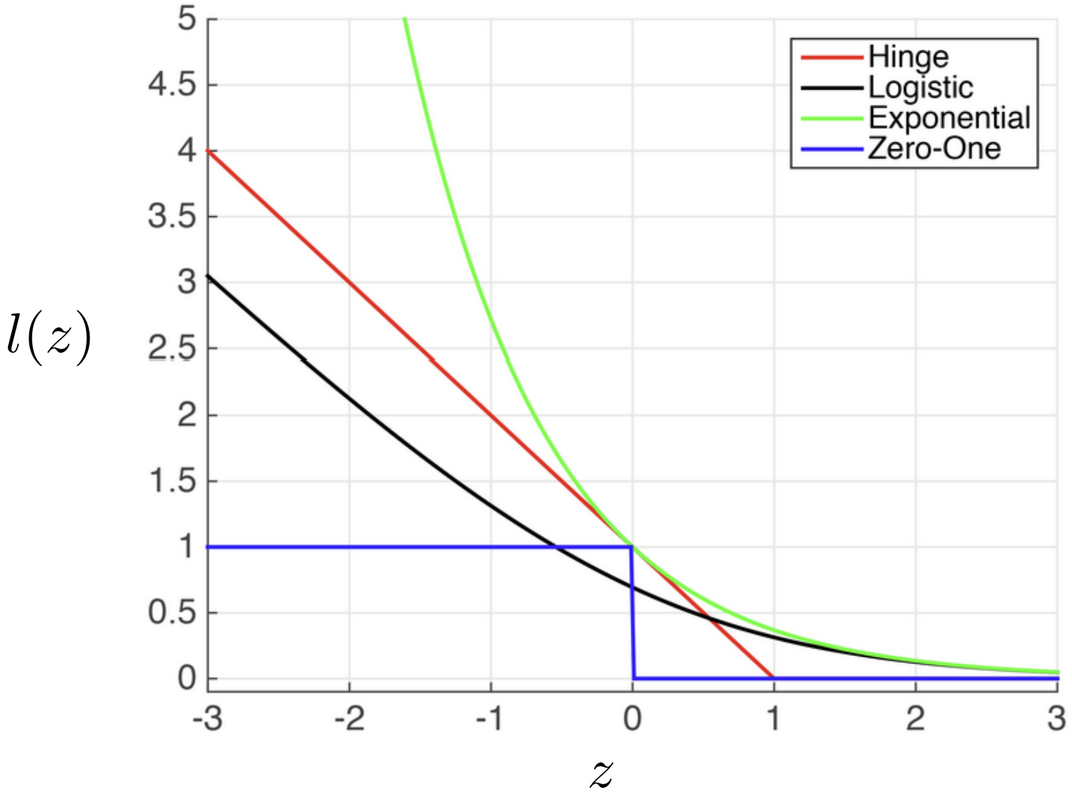
In logistic regression look for $\beta \in \mathbb{R}^p$ that minimizes:

$$-\sum_{y_i=1} \log\left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right) - \sum_{y_i=-1} \log\left(1 - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)$$

- $\log\left(1 - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right) = -l(\beta_i^x) = -l(\beta^T x_i \cdot y_i)$ if $y_i = 1$
- $\log\left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right) = \log\left(\frac{1}{e^{-\beta^T x_i}+1}\right) = -l(\beta^T x_i \cdot y_i)$ if $y_i = -1$.

Classification losses $Y \in \{-1, 1\}$

Minimize $\frac{1}{n}\sum_{i=1}^{n} l(h_w(x_i) \cdot y_i) + \lambda r(y_i)$ $\quad w \in \mathcal{R}^q$ $\quad h_w(x_i)y_i \mapsto z$

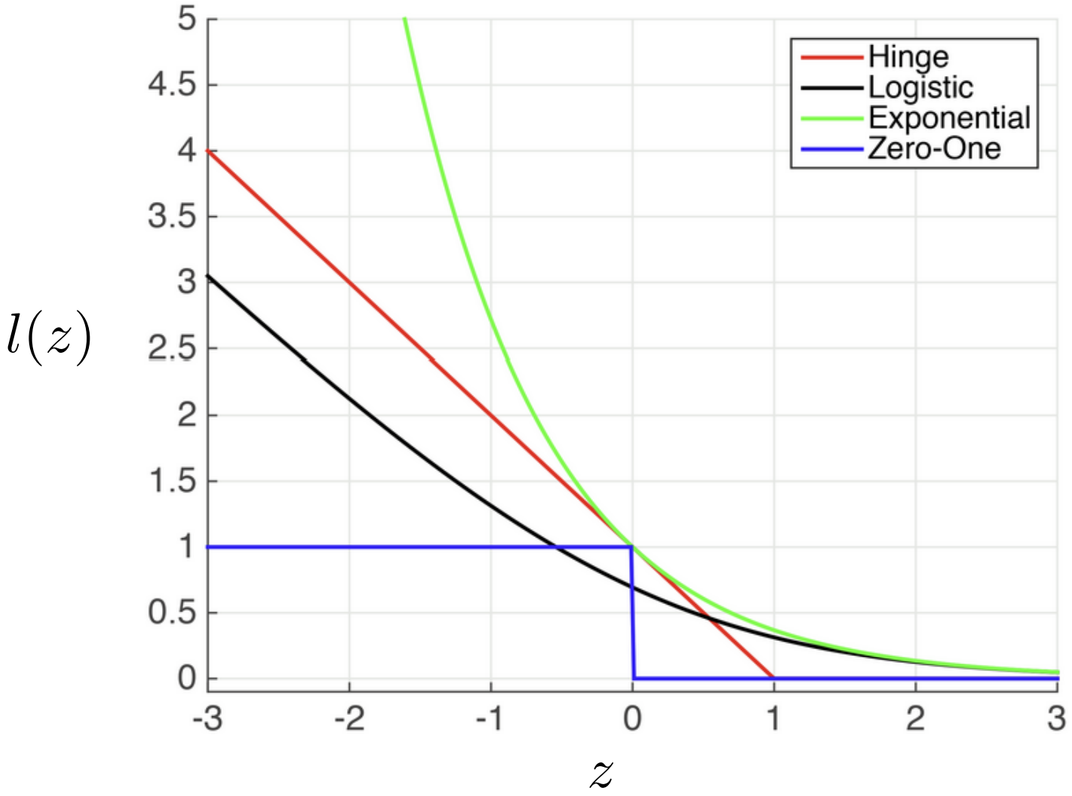


Logistic loss: $l(z) = \log(1 + e^{-z})$
Used for logistic regression, probabilistic interpretation

Hinge loss: $l(z) = \max(1 - z, 0)$
Used for Support vector machines (distance between margin and closest point)

Classification losses $Y \in \{-1, 1\}$

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} l(h_w(x_i) \cdot y_i) + \lambda r(y_i) \qquad w \in \mathcal{R}^q \qquad h_w(x_i) y_i \mapsto z$$



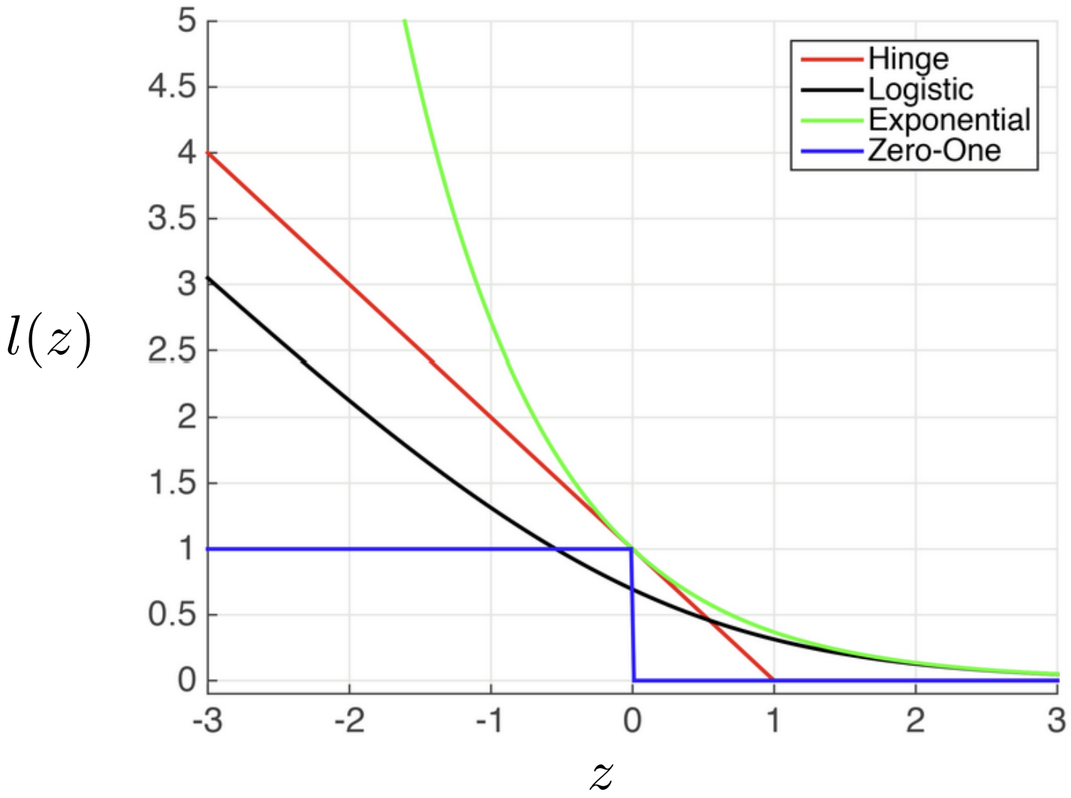
Logistic loss: $l(z) = \log(1 + e^{-z})$
Used for logistic regression, probabilistic interpretation

Hinge loss: $l(z) = \max(1 - z, 0)$
Used for Support vector machines (distance between margin and closest point)

Exponential loss: $l(z) = e^{-z}$
Agressive loss used for Adaboost (very sensitive to noise, works in very particular cases)

Classification losses $Y \in \{-1, 1\}$

$$\text{Minimize} \frac{1}{n} \sum_{i=1}^{n} l(h_w(x_i) \cdot y_i) + \lambda r(y_i) \qquad w \in \mathcal{R}^q \qquad h_w(x_i) y_i \mapsto z$$



Logistic loss: $l(z) = \log(1 + e^{-z})$
Used for logistic regression, probabilistic interpretation

Hinge loss: $l(z) = \max(1 - z, 0)$
Used for Support vector machines (distance between margin and closest point)

Exponential loss: $l(z) = e^{-z}$
Agressive loss used for Adaboost (very sensitive to noise, works in very particular cases)

Zero-one loss: $l(z) = 1_{z<0}$
Final loss used to evaluate the performance of a model. Not continuous so almost impractible for optimization