# Introduction to Statistical Learning Theory

March 10, 2023

# Statistics vs Learning

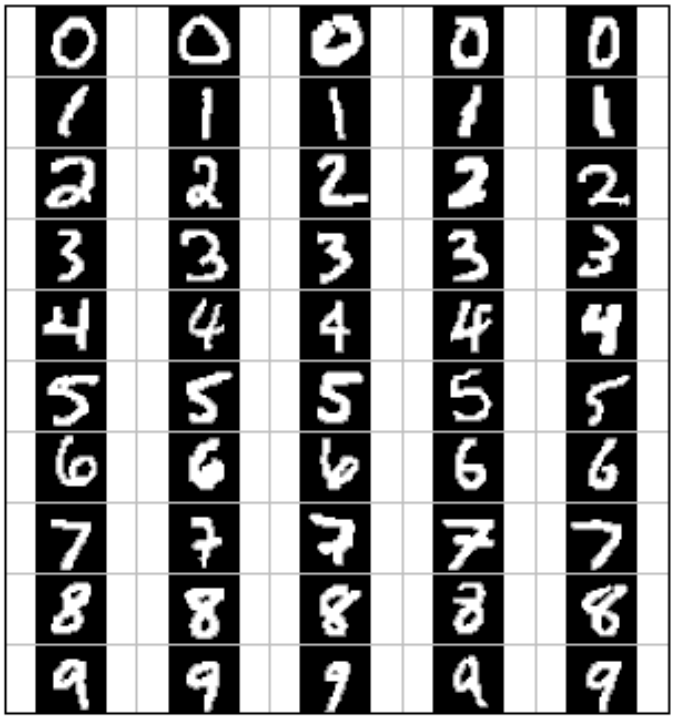
**Statistics:**
Observations $(X_i)_{1 \leq i \leq n}$
Model selection $(Z, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$.
$\mapsto$ Estimation of $\theta$
$\mapsto$ Test
Evaluation of the overall procedure on $\theta$.

**Learning:**
Observations $(X_i, Y_i)_{1 \leq i \leq n}$
Goal: **forecast** $Y$ as a funtion of $X$.
$\mapsto$ **Approach 1:** Define a model on the joint law of $(X, Y)$, learn the parameters and deduce a decision rule
$\mapsto$ **Approach 2:** Choose a class of functions, a loss measurement and find the best rule in the class according to the risk measure
Evaluation through a risk on the classification rule

# Example: Spam detection

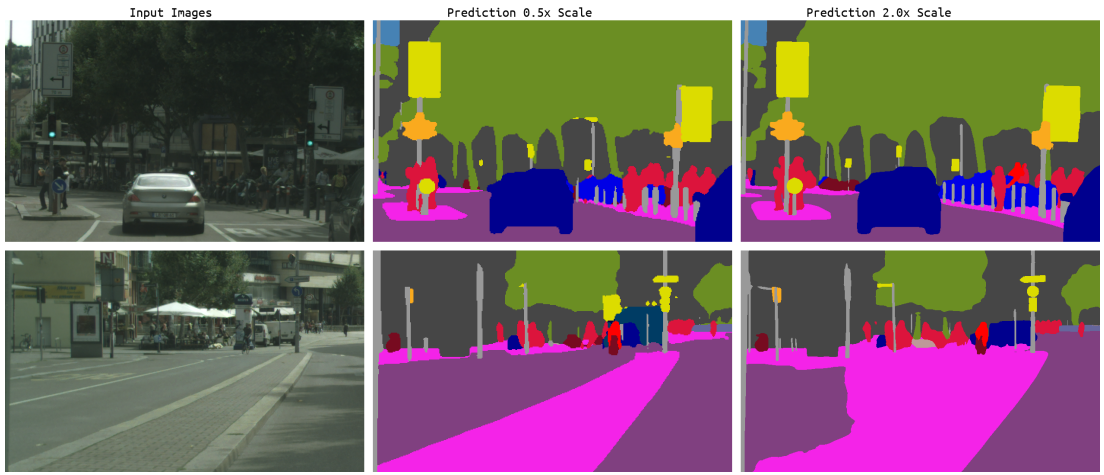

- Dataset: e-mails
- Input: (texts, headers - senders, receivers, smtp, ...)
- Output : "Spam"/"Non Spam"

# Handwritten recognition



- Dataset: images of numbers (each image is a $28 \times 28 = 784$ dimensional vector - greyscale level)
- Input: Image
- Output: recognized number

# Image segmentation




- Learning set: annotated and segmented images
- Output: object

# 2 examples with a univariate input $X$ $(d = 1)$

Height & Gender:

Age & Cardiovascular disease:

# Supervised learning

## Supervised learning

- Inputs: $\mathbf{X} \in \mathsf{X} = \mathbb{R}^d$
- Outputs: $Y \in \mathsf{Y}$.
- $(\mathbf{X}, Y) \sim \mathbb{P}$ where $\mathbb{P}$ is unknown.
  - $Y \in \{0, 1\}$ (or $\{-1, 1\}$) - binary classification,
  - $Y \in \{0, 1, 2, \ldots, K-1\}$ - multiclass problem
  - $Y \in \mathbb{R}$ - regression.

- A classifier or predictor is a function $g : \mathsf{X} \to \mathsf{Y}$ measurable.

## Goal

- Learning set : $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbb{P}$)
- Build a predictor / classifier $\widehat{g}_n$ from $\mathcal{D}_n$

# Loss and risk

## Loss function

- Loss function : $\ell(y, g(\mathbf{x}))$ quantifies the quality of forecasting of $y$ with $g(\mathbf{x})$.
- Examples:
  - 0-1 loss (classification): $\ell(y, g(\mathbf{x})) = \mathbb{1}_{\{y \neq g(\mathbf{x})\}}, \quad y \in \mathsf{Y} = \{0, 1\}$
  - Quadratic loss (regression): $\ell(y, g(\mathbf{x})) = \|y - g(\mathbf{x})\|^2$, $y \imath \mathbb{R}^d$

## Risk of a decision rule

- Risk: $R(g) = \mathbb{E}[\ell(Y, g(\mathbf{X}))] = \int \ell(y, g(\mathbf{x})) \mathrm{d}\mathbb{P}(x, y)$
- Examples:
  - 0-1 risk (classification) : $R(g) = \mathbb{P}(Y \neq g(\mathbf{X}))$
  - Quadratic risk (regression): $R(g) = \mathbb{E}[\|Y - g(\mathbf{X})\|^2]$

# Supervised learning: summary

- Learning set : $D_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbb{P}$)
- Decision rule (classifier, predictor): $g : \mathsf{X} \to \mathsf{Y}$ measurable
- Loss: $\ell(Y, g(\mathbf{X}))$
- Risk: $R(g) = \mathbb{E}[\ell(Y, g(\mathbf{X}))] = \int \ell(y, g(\mathbf{x}))\mathrm{d}\mathbb{P}(\mathbf{x}, y)$

## Questions ?

- Existence of a decision rule $g^*$ optimal, $R(g^*) \leq R(g)$ for any other decision rule $g$ ?
- How to build $\widehat{g}_n$ ?
- Can we assess a result on $R(\widehat{g}_n) - R(g^*)$ ?

# Bayes risk / Bayes Classifier

## Definition

Let $\ell : \mathsf{Y} \times \mathsf{Y} \to [0, \infty)$ a loss function, $\mathbb{P}$ a probability distribution on $\mathsf{X} \times \mathsf{Y}$. The minimum risk

$$R^*_{\ell,\mathbb{P}} := \inf \left\{ R_{\ell,\mathbb{P}}(g) = \int \ell(y, g(\mathbf{x})) \mathrm{d}\mathbb{P}(\mathbf{x}, y) \ , g : \mathsf{X} \to \mathsf{Y} \text{ measurable} \right\}$$

is the Bayes risk relative to $\mathbb{P}$ and $\ell$. A measurable function $g^*_{\ell,\mathbb{P}} : \mathsf{X} \to \mathsf{Y}$ that satisfies

$$R_{\ell,\mathbb{P}}\left(g^*_{\ell,\mathbb{P}}\right) = R^*_{\ell,\mathbb{P}}$$

is a Bayes decision rule

For the sake of simplicity we will denote $R^*_{\ell,\mathbb{P}} = R^*$

# Conditional risk

$$R_{\ell,\mathbb{P}}(g) \stackrel{\text{def}}{=} \mathbb{E}\left[\ell(Y, g(\mathbf{X}))\right] = \mathbb{E}\left[\mathbb{E}\left[\ell(Y, g(\mathbf{X})) \,|\, \mathbf{X}\right]\right] = \int R_{\ell,\mathbb{P}}(g, \mathbf{x}) \mathrm{d}\mathbb{P}(\mathbf{x})$$

where $R_{\ell,\mathbb{P}}(g, \mathbf{x})$ is the conditional risk:

$$R_{\ell,\mathbb{P}}(g, \mathbf{x}) = \int \ell(y, g(\mathbf{x})) \mathrm{d}\mathbb{P}_{Y|\mathbf{X}}(y \mid \mathbf{x})$$

The Bayesian decision rule is obtained, for any $\mathbf{x} \in \mathsf{X}$, while minimizing the conditional risk, i.e.

$$R_{\ell,\mathbb{P}}(g^*_{\ell,\mathbb{P}}(\mathbf{x}), \mathbf{x}) \leq R_{\ell,\mathbb{P}}(g, \mathbf{x})$$

so that:

$$g^*_{\ell,\mathbb{P}}(\mathbf{x}) = \inf_a \int \ell(y, a) \mathrm{d}\mathbb{P}_{Y|\mathbf{X}}(y \mid \mathbf{x})$$

We skip some technical details (including the definition of the conditional distribution of $Y|X$).

# 1. Regression: Bayesian predictor

**Theorem**

*For any predictor* $g : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}[(Y - g^*(\mathbf{X}))^2] \leq \mathbb{E}[(Y - g(\mathbf{X}))^2]$$

*where* $g^*(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$

**Proof.**

Definition of the conditional expectation! □

Density case: we compute

$$g^*(\mathbf{x}) = \int_{\mathbb{R}^d} y p_{Y|\mathbf{X}}(y|\mathbf{x}) \mathrm{d}y,$$

with

$$p_{Y|\mathbf{X}}(y|\mathbf{x}) = \frac{p_{Y,\mathbf{X}}(y, \mathbf{x})}{p_{\mathbf{X}}(\mathbf{x})}$$

.

- label: $(\mathsf{Y}, \mathcal{Y}, \nu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathrm{Leb})$, Covariates: $(\mathsf{X}, \mathcal{X}, \mu) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathrm{Leb})$
- $q_{\mathbf{X}|Y}(\mathbf{x}|\mathbf{y})$ conditional density of the covariates $\mathbf{X}$ given the output $Y$
- Marginal density of $Y$: $p_Y(y)$

## Joint distribution

$$p_{Y,\mathbf{X}}(y, \mathbf{x}) = q_{\mathbf{X}|Y}(\mathbf{x}|y) p_Y(y)$$

## Marginal distribution of the covariates

$$p_{\mathbf{X}}(\mathbf{x}) = \int_{\mathsf{Y}} p_{Y,\mathbf{X}}(y, \mathbf{x}) \mathrm{d}y$$

## Conditional distribution of the output given the covariates

$$p_{Y|\mathbf{X}}(y|\mathbf{x}) = \frac{p_{Y,\mathbf{X}}(y, \mathbf{x})}{p_{\mathbf{X}}(\mathbf{x})} = \frac{q_{\mathbf{X}|Y}(\mathbf{x}|y) p(y)}{\int_{\mathsf{Y}} q_{\mathbf{X}|Y}(\mathbf{x}|y') p_Y(y') \mathrm{d}y'}$$

# 2. Classification - Bayesian classifier

Framework: $Y \in \{0, 1\}$.
Loss: $\ell_{01}(y, g(\mathbf{x})) = \mathbb{1}_{\{y \neq g(\mathbf{x})\}}$, $y \in \mathsf{Y} = \{0, 1\}$

We define $\eta(\mathbf{X}) = \mathbb{E}[Y \,|\, \mathbf{X}] = \mathbb{P}(Y = 1 \,|\, \mathbf{X})$ as the *regression function*.

**Theorem (Bayesien classifier)**

*For any classifier* $g : \mathbb{R}^d \to \{0, 1\}$,

$$\mathbb{E}[\ell_{01}(Y, g^*(\mathbf{X}))] \leq \mathbb{E}[\ell_{01}(Y, g(\mathbf{X}))]$$

*where* $g^*(\mathbf{X}) = \mathbb{1}_{\{\eta(\mathbf{X}) \geq 1/2\}}$.

- It essentially corresponds to return the most likely label when seeing $\mathbf{X}$.

**Theorem (Excess risk)**

*For any* $g$ *the excess risk satisfies:*

$$R_{01}(g) - R_{01}(g^*) = \mathbb{E}\left[|2\eta(\mathbf{X}) - 1| \mathbb{1}_{\{g^*(\mathbf{X}) \neq g(\mathbf{X})\}}\right]$$

# Proof

# Explicitation of the Bayesian rule

- Label: $(\mathsf{Y}, \mathcal{Y}, \nu) = (\{0,1\}, \mathcal{P}(\{0,1\}), \mathrm{c})$ where $\mathrm{c} = \delta_{\{0\}} + \delta_{\{1\}}$ is the counting measure. We denote $\mathcal{B}(p)$ the Bernoulli distribution of parameter $p \in [0,1]$,

$$\mathcal{L}(Y) = \mathcal{B}(p)$$

- Covariates: $(\mathsf{X}, \mathcal{X}, \mu) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathrm{Leb})$
- $q_0, q_1$ densities of the covariates given the class $0, 1$

$$\mathcal{L}(\mathbf{X} \,|Y = 0) = q_0 \mathrm{Leb} \qquad \mathcal{L}(\mathbf{X} \,|Y = 1) = q_1 \mathrm{Leb}$$

- Marginal distribution of the labels $y \mapsto p\mathbb{1}_{\{1\}}(y) + (1-p)\mathbb{1}_{\{0\}}(y)$

**Joint distribution of the labels and covariates**

$$p_{Y,\mathbf{X}}(y, \mathbf{x}) = pq_1(\mathbf{x})\mathbb{1}_{\{1\}}(y) + (1-p)q_0(\mathbf{x})\mathbb{1}_{\{0\}}(y)$$

# Regression function

## Marginal distribution of the covariates

$$p_{\mathbf{X}}(\mathbf{x}) = \int_{\mathsf{Y}} p_{Y,\mathbf{X}}(y,\mathbf{x})\,\mathrm{c}(\mathrm{d}y) = pq_1(\mathbf{x}) + (1-p)q_0(\mathbf{x})$$

## Marginal distribution of the labels given the covariates

$$\begin{cases} p_{Y|\mathbf{X}}(1|\mathbf{x}) &= \frac{p_{Y,\mathbf{X}}(1,\mathbf{x})}{p_{\mathbf{X}}(\mathbf{x})} = \frac{pq_1(\mathbf{x})}{pq_1(\mathbf{x})+(1-p)q_0(\mathbf{x})} \\ p_{Y|\mathbf{X}}(0|\mathbf{x}) &= \frac{p_{Y,\mathbf{X}}(0,\mathbf{x})}{p_{\mathbf{X}}(\mathbf{x})} = \frac{(1-p)q_0(\mathbf{x})}{pq_1(\mathbf{x})+(1-p)q_0(\mathbf{x})} \end{cases}$$

## Regression function

$$\eta(\mathbf{x}) := p_{Y|\mathbf{X}}(1|\mathbf{x}) = \frac{pq_1(\mathbf{x})}{pq_1(\mathbf{x}) + (1-p)q_0(\mathbf{x})}\ .$$

# Bayesian classifier

Recall that:

$$g^*(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad \eta(\mathbf{x}) \geq 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

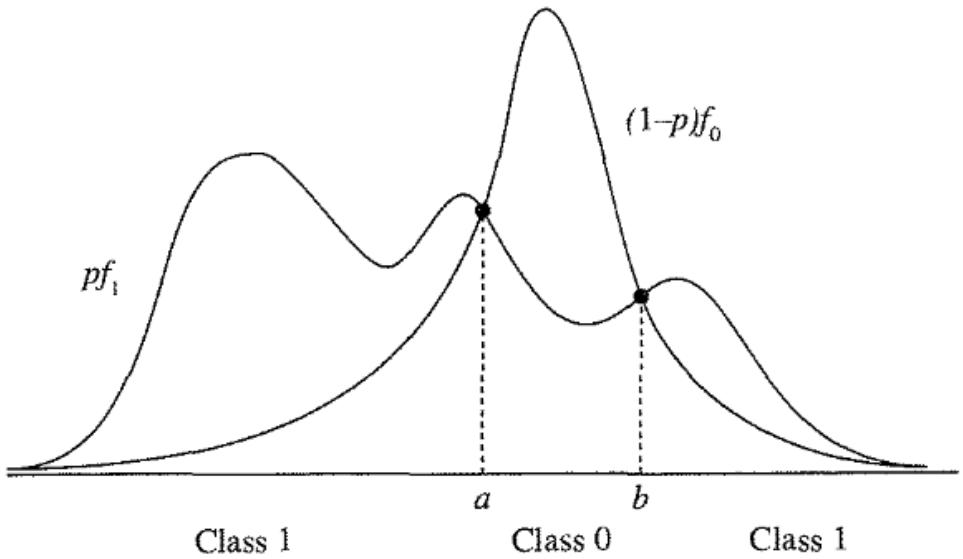



Figure: Example of the Bayes decision rule: $pf_1/(pf_1 + (1-p)f_0)$ is the probability of $\mathbf{Y} = 1$ given $\mathbf{X}$

# Summary: Bayesian classifier

- Problem: In practice, we do not know the joint distribution $(\mathcal{L}(\mathbf{X}, Y))$.
- *Theoretical fact:* the Bayesian classifier is optimal but is not a function of the observations.
- *Solution* Learning set : $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbb{P}$)
- Bayesian error is solely a theoretical object we will compare with!
- There are two random sources: the one of the *new* $\mathbf{X}, Y$ and the one brought by $\mathcal{D}_n$.

# Leading idea

1. We **model** $(\mathbf{X}, Y)$ with a (parametric): $\{\mathbb{P}_\theta, \theta \in \Theta\}$
2. We compute $g^*_\theta$ under any $\mathbb{P}_\theta$.
3. We estimate $\theta$ with a statistical method $\hat{\theta}_n$.
4. We use the plug-in rule $\hat{g}_n = g^*_{\hat{\theta}_n}$ (plug-in $\hat{\theta}_n$ in $\theta \mapsto g^*_\theta$).

Remark : model the joint law $(\mathbf{X}, Y)$

# Leading idea

1. We **model** $(\mathbf{X}, Y)$ with a (parametric): $\{\mathbb{P}_\theta, \theta \in \Theta\}$
2. We compute $g^*_\theta$ under any $\mathbb{P}_\theta$.
3. We estimate $\theta$ with a statistical method $\hat{\theta}_n$.
4. We use the plug-in rule $\hat{g}_n = g^*_{\hat{\theta}_n}$ (plug-in $\hat{\theta}_n$ in $\theta \mapsto g^*_\theta$).

Remark : model the joint law $(\mathbf{X}, Y)$

- Solution 1: model $Y|\mathbf{X}$ and $\mathbf{X}$ (logistic regression, KNN, linear regression)
- Solution 2: model $\mathbf{X}|Y$ and $Y$ (LDA, QDA, Naive Bayes).

# Discriminant analysis:

- Model the conditional distribution of $\mathbf{X}|Y$ for any value of $Y$ with a Gaussian distribution

$$q_i \in \left\{ \mathrm{N}(\mu_i, \Gamma_i) \ : \ \mu_i \in \mathbb{R}^d, \Gamma_i \in \mathrm{S}_d^+ \right\}$$

  where $\mathrm{S}_d^+$ is the set of symetric positive definite matrices.
- Regression function:

$$\eta(\mathbf{x}) = \frac{p q_1(\mathbf{x})}{p q_1(\mathbf{x}) + (1-p) q_0(\mathbf{x})}$$

- Classification rule:

$$\eta(\mathbf{x}) \geq 1/2 \iff p q_1(\mathbf{x}) \geq (1-p) q_0(\mathbf{x})$$

  or in a similar way:

$$\log q_1(\mathbf{x}) - \log q_0(\mathbf{x}) + \log\left(\frac{p}{1-p}\right) \geq 0 \ .$$

# Discriminant analysis and Gaussian modelling: practical setup

- Gaussian case,

$$\log q_1(\mathbf{x}) - \log q_0(\mathbf{x}) = \frac{1}{2}\log\left(\frac{|\Gamma_0|}{|\Gamma_1|}\right) + \log\left(\frac{p}{1-p}\right) + \frac{1}{2}(\mathbf{x}-\mu_0)^T\Gamma_0^{-1}(\mathbf{x}-\mu_0) - \frac{1}{2}(\mathbf{x}-\mu_1)^T\Gamma_1^{-1}(\mathbf{x}-\mu_1)$$

- If $\Gamma_0 = \Gamma_1$ [same intra-class covariance matrix], the decision rule is linear (see tutorials): LDA: Linear Discriminant Analysis.
- If $\Gamma_0 \neq \Gamma_1$, the decision rule is quadratic: QDA (Quadratic Discriminant Analysis).

# Parameters estimation and plug-in.

- For any class, estimation of the mean and covariance [for example, MLE]

$$\hat{\mu}_i = \frac{\sum_{k=1}^n \mathbb{1}_{\{i\}}(Y_k)\mathbf{X}_k}{\sum_{k=1}^n \mathbb{1}_{\{i\}}(Y_k)} \qquad i \in \{0,1\}$$

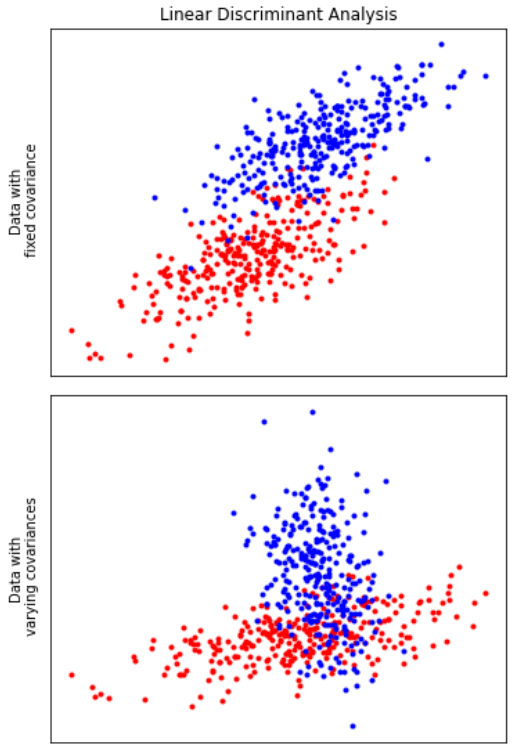
$$\hat{\Gamma}_i = \frac{\sum_{k=1}^n \mathbb{1}_{\{i\}}(Y_k)(\mathbf{X}_k - \hat{\mu}_i)(\mathbf{X}_k - \hat{\mu}_i)^T}{\sum_{k=1}^n \mathbb{1}_{\{i\}}(Y_i)}$$

$$\hat{p} = n^{-1} \sum_{i=1}^n \mathbb{1}_{\{1\}}(Y_i)$$

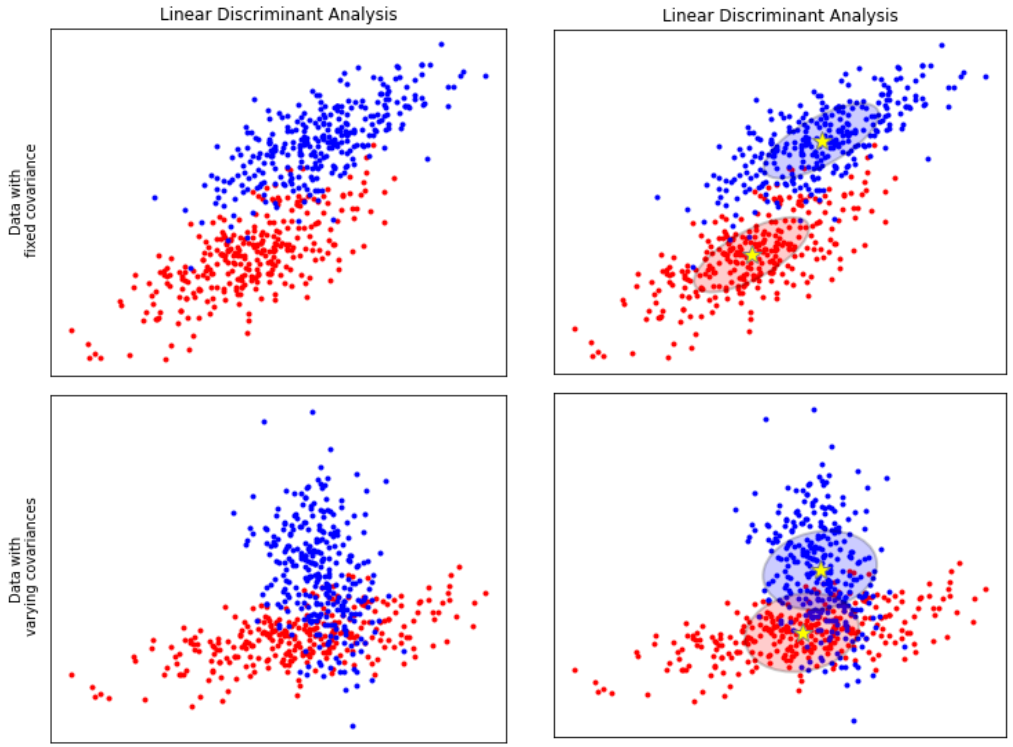
- Plug-in: Decision rule

$$\frac{1}{2}\log\left(\frac{|\hat{\Gamma}_0|}{|\hat{\Gamma}_1|}\right) + \log\left(\frac{\hat{p}}{1-\hat{p}}\right)$$
$$+ \frac{1}{2}(\mathbf{x}-\hat{\mu}_0)^T\hat{\Gamma}_0^{-1}(\mathbf{x}-\hat{\mu}_0) - \frac{1}{2}(\mathbf{x}-\hat{\mu}_1)^T\hat{\Gamma}_1^{-1}(\mathbf{x}-\hat{\mu}_1) \geq 0$$

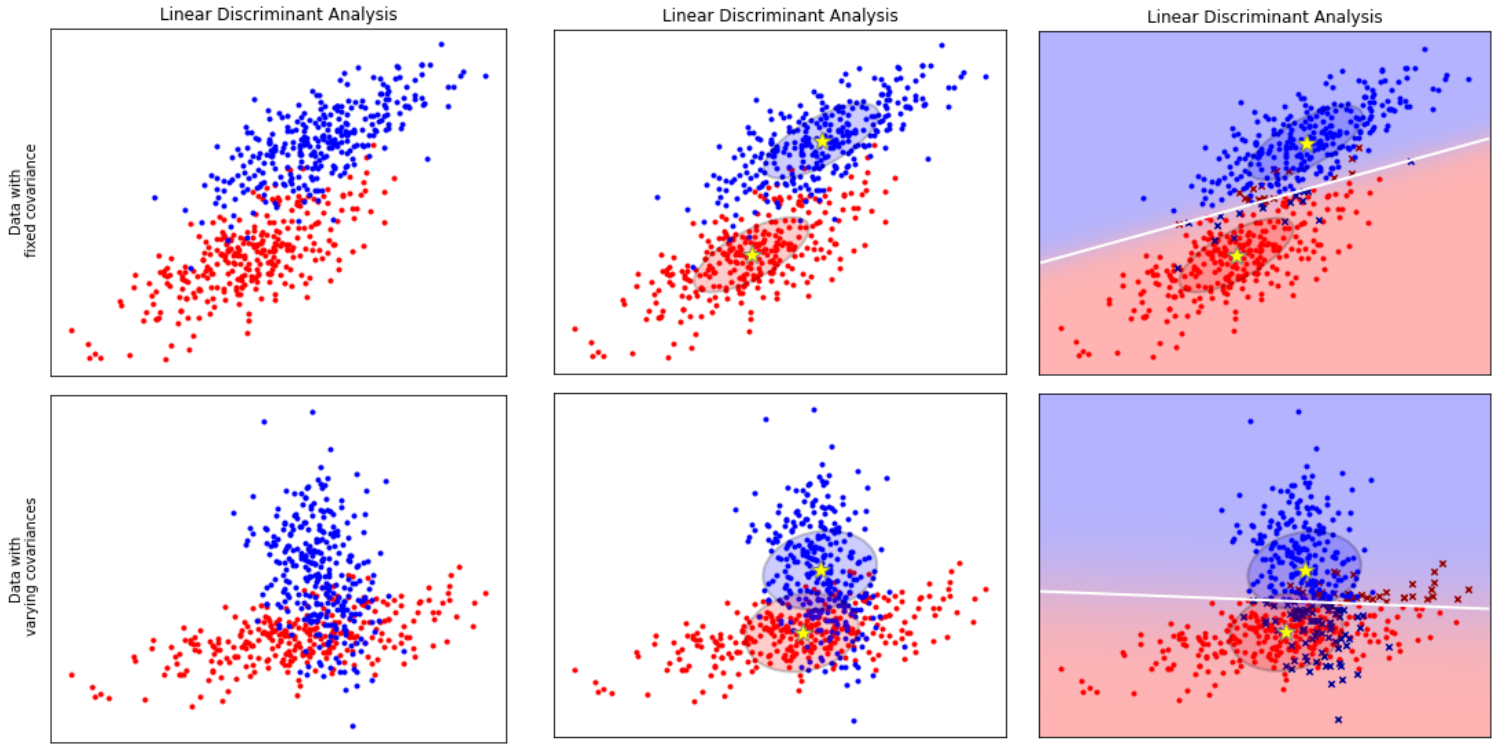
# Example: LDA


Linear Discriminant Analysis
Data with fixed covariance
Data with varying covariances

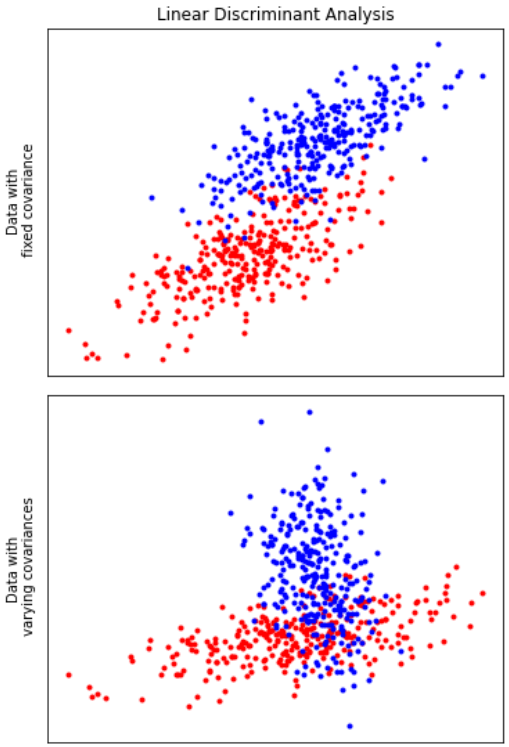
# Example: LDA


Linear Discriminant Analysis
Linear Discriminant Analysis
Data with fixed covariance
Data with varying covariances

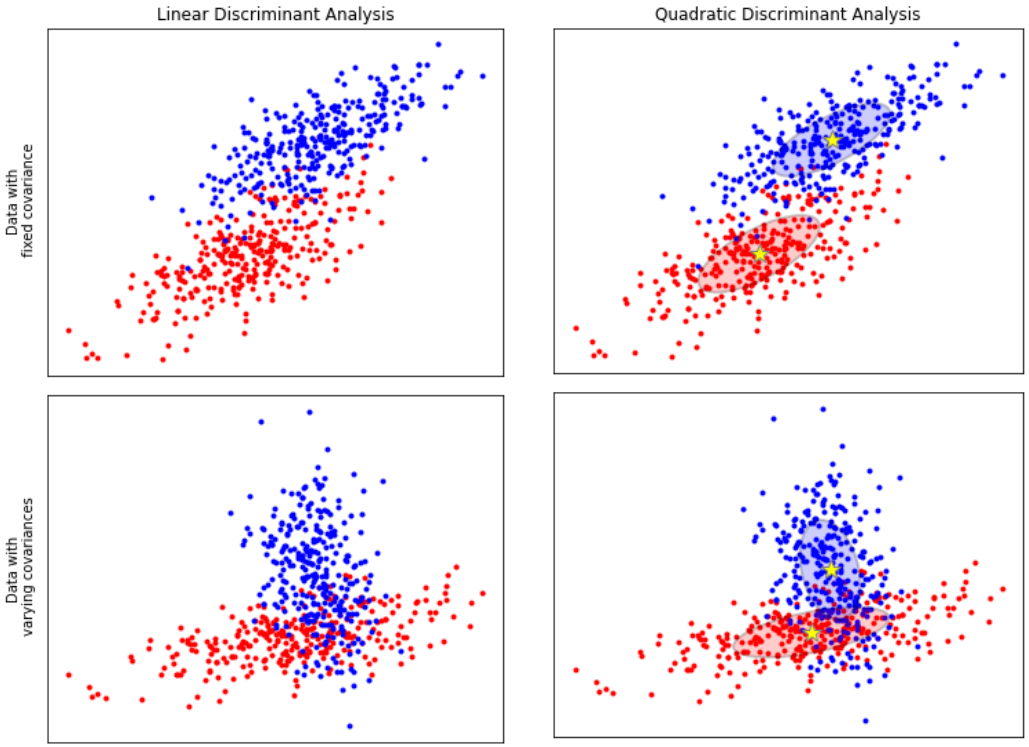
# Example: LDA


Linear Discriminant Analysis
Linear Discriminant Analysis
Linear Discriminant Analysis
Data with fixed covariance
Data with varying covariances

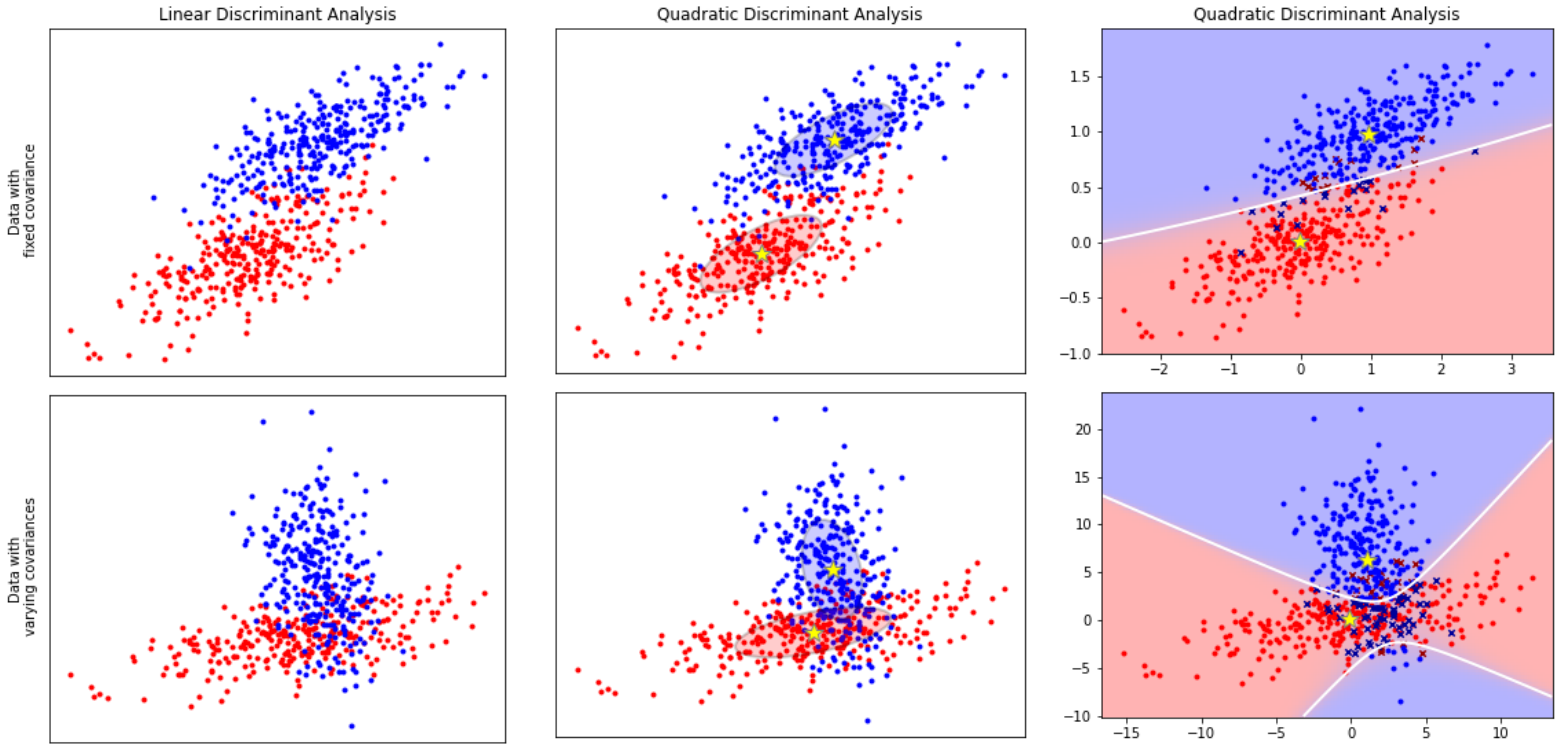
# Example: QDA


Linear Discriminant Analysis
Data with fixed covariance
Data with varying covariances

# Example: QDA


Linear Discriminant Analysis
Quadratic Discriminant Analysis
Data with fixed covariance
Data with varying covariances

# Example: QDA


Linear Discriminant Analysis
Quadratic Discriminant Analysis
Quadratic Discriminant Analysis
Data with fixed covariance
Data with varying covariances

# Naive Bayes method

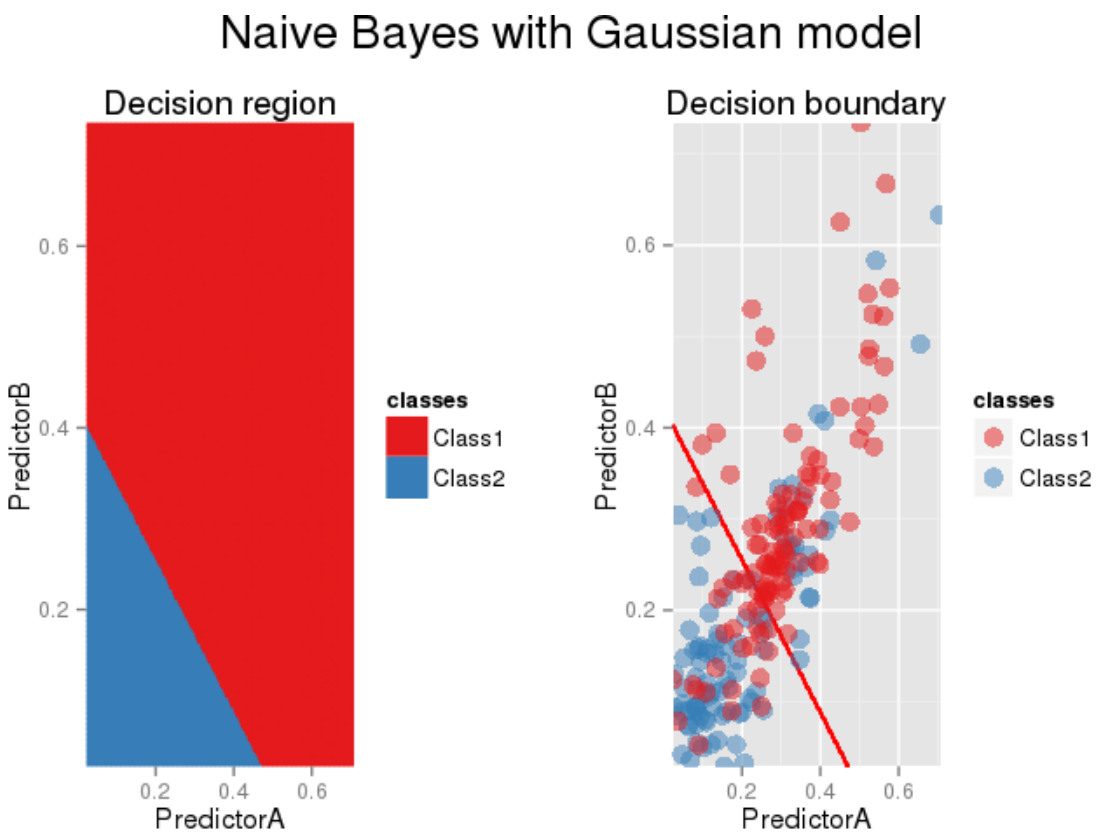
## Naive Bayes

- The conditional distributions are assumed to be *independent*

$$q_i(\mathbf{x}) = \prod_{j=1}^{d} \tilde{q}_{i,j}(x^{(j)}) \,, \quad \mathbf{x} = (x^{(1)}, \ldots, x^{(d)})$$

- Example:
  - Numerical variable: $\tilde{q}_{i,j}$ with mean $\mu_{i,j}$ and variance $\sigma^2_{i,j}$
  - Categorical variable: binomial or multinomial distribution.
- If all numerical variables are Gaussian, approach equivalent to LDA or QDA with diagonal covariance matrices
- Advantage Simple to set up even in very large dimensions!

# Example: Naive Bayes



# Example : Naive Bayes

- For numerical variables, we can replace the Gaussian model by more sophisticated density models like mixture models

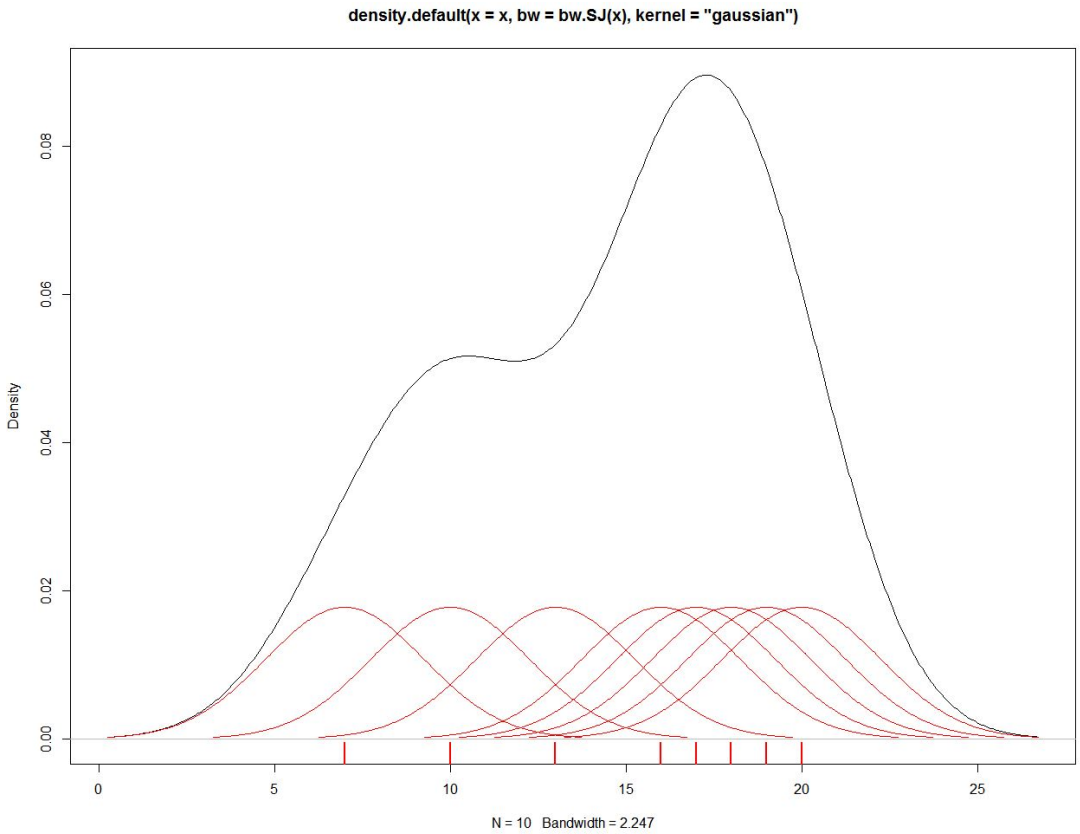
$$q_{i,j}(x^{(j)}) = \sum_{j=1}^{r} \pi_{i,j,r} \, \mathrm{N}(\mu_{i,j,r}, \sigma^2_{i,j,r})$$

- It is also possible to use non-parametric statistics to estimate $q_{i,j}$, with kernel estimators

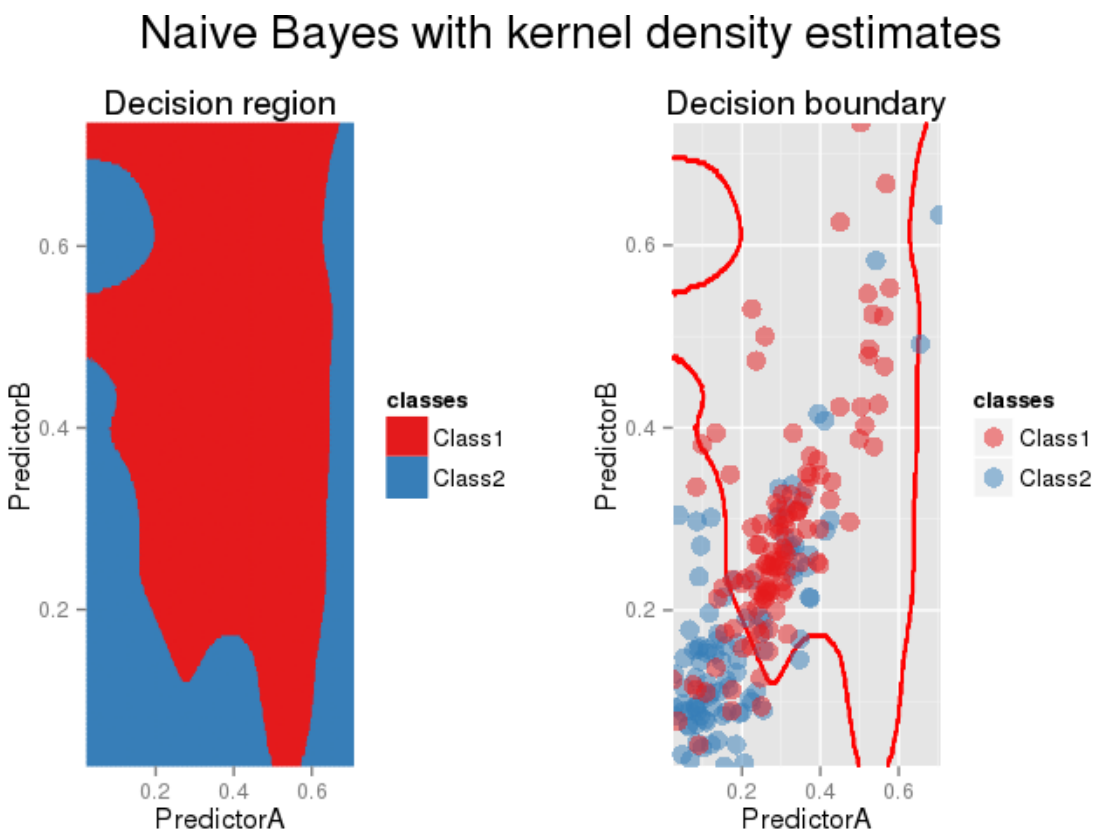
$$\hat{q}_{i,j}(x^{(j)}) = \frac{\sum_{\ell=1}^{n} \mathbb{1}_{\{i\}}(Y_\ell) K(x^{(j)} - \mathbf{X}_\ell^{(j)})}{\sum_{\ell=1}^{n} \mathbb{1}_{\{i\}}(Y_\ell)}$$

where $K$ is a kernel, $\int K(x)\mathrm{d}x = 1$, $K(x) = K(-x)$

# Example: Naive Bayes



# Example: Naive Bayes



Goal: learn a decision rule $\widehat{g}$ from the learning set $D_n$

## $\mapsto$ Generative approache

Solution: Estimate the regression function $\eta(X) = \mathbb{P}(Y = 1 \,|\, X)$ and substitute (plug-in) this estimator inside the Bayes' rule: Linear models, kernel methods, $k$-nearest neighbors, Naive Bayes,...

## $\mapsto$ Optimization approach

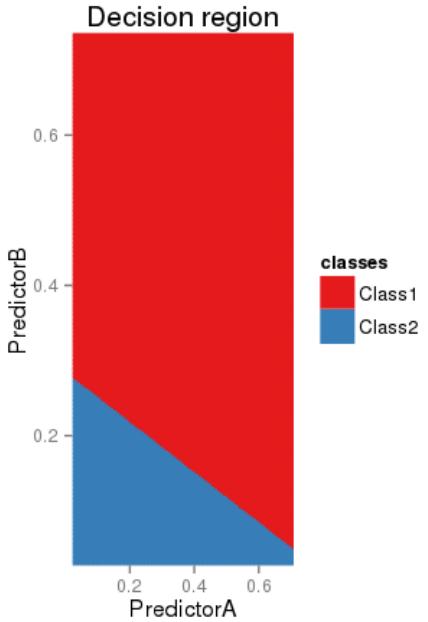
Solution: Minimize the empirical risk (or an upper bound of the empirical risk): support vector machines, boosting, neural networks,...

## Canonical example

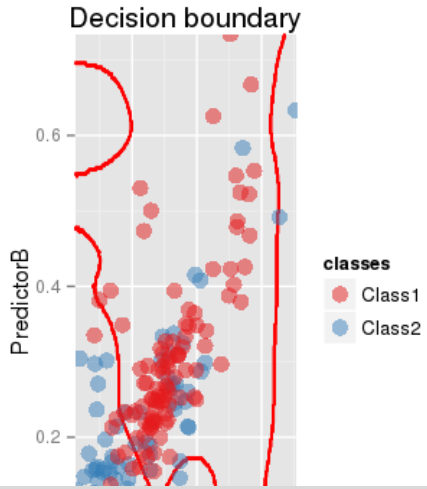
- Restrict the decision rules $g$ to the set of linear separations $\mathcal{G}$
  - Linear classification: $\mathcal{G} = \left\{ g_{(b,\mathbf{w})}(\mathbf{x}) = \mathbb{1}_{b+\mathbf{w}^T\mathbf{x} \geq 0} \ : \ (b, \mathbf{w}) \in \mathbb{R} \times \mathbb{R}^d \right\}$
- Minimize the empirical risk:

$$\widehat{g} = \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, g(\mathbf{X}_i))$$

# Example: linear discrimination (left) or more complex class (right)

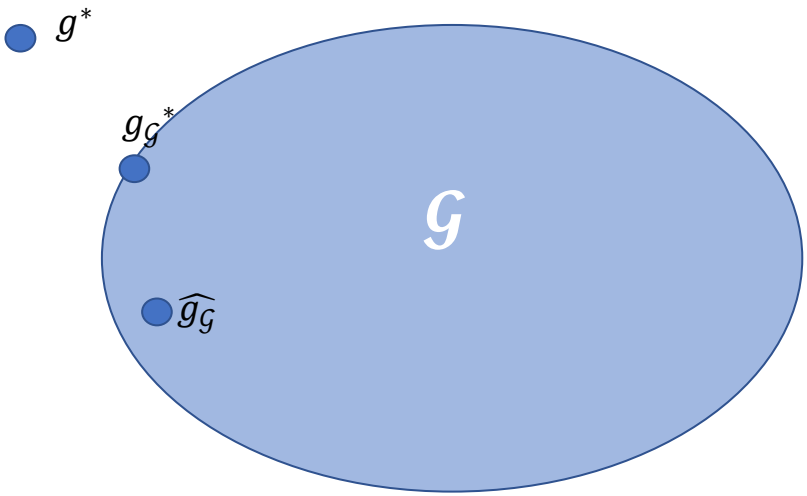




# Risk decomposition

- General framework:
  - $\mathcal{G}$: classification rules
  - Optimal classifier: $g^*$ Bayes' rule
  - Optimal classifier in $\mathcal{G}$: $g_{\mathcal{G}}^* = \arg\min_{g \in \mathcal{G}} R(g)$




**Risk decomposition**

$$R(\widehat{g}_{\mathcal{G}}) - R(g^*) = \underbrace{R(g_{\mathcal{G}}^*) - R(g^*)}_{\text{Approximation error}} + \underbrace{R(\widehat{g}_{\mathcal{G}}) - R(g_{\mathcal{G}}^*)}_{\text{Estimation error}}$$

- The approximation error may be large if $\mathcal{G}$ is not well chosen
- The estimation error may be large if $\mathcal{G}$ is too complex (to be properly defined later on).

$\mapsto$ Next slides: upper bound the estimation error!

# Over-fitting / Under-fitting

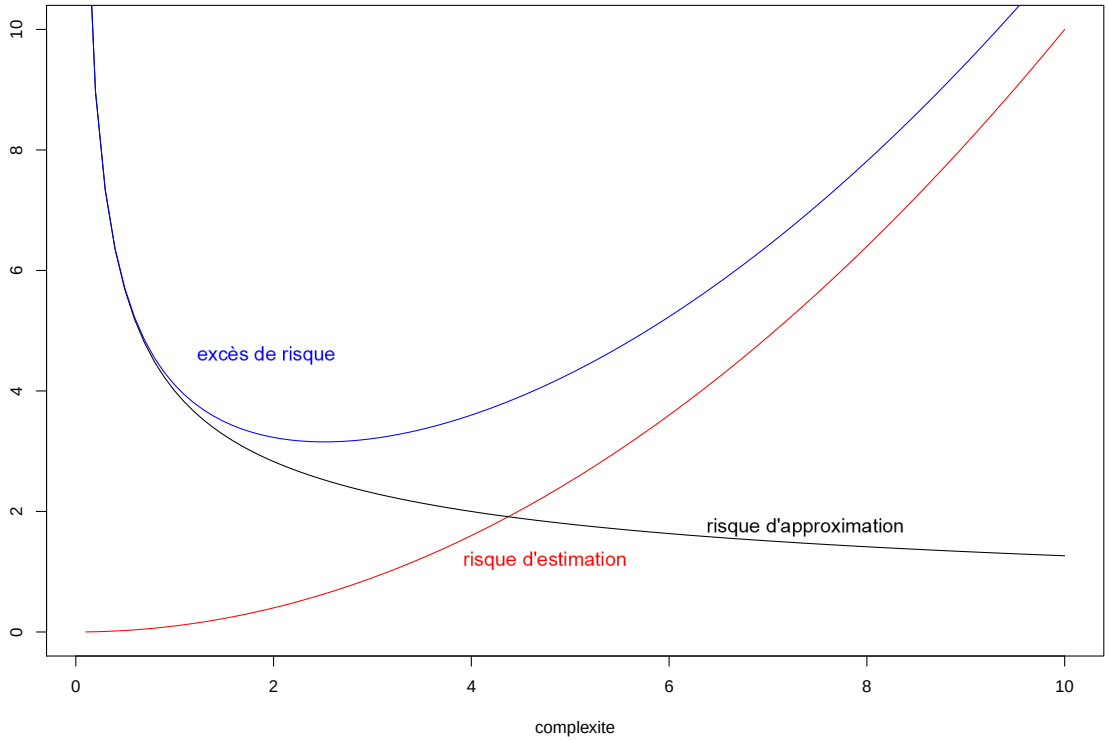
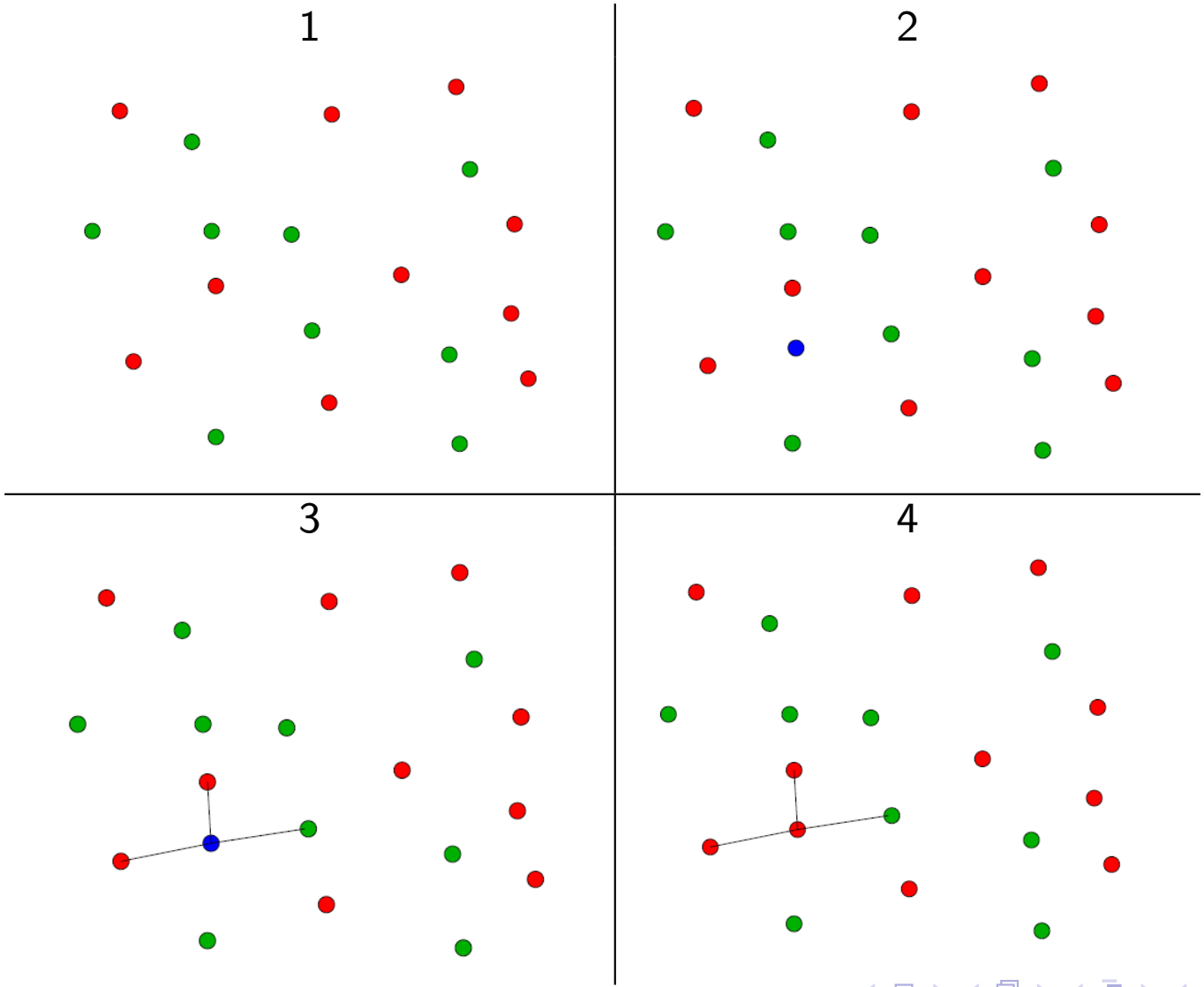
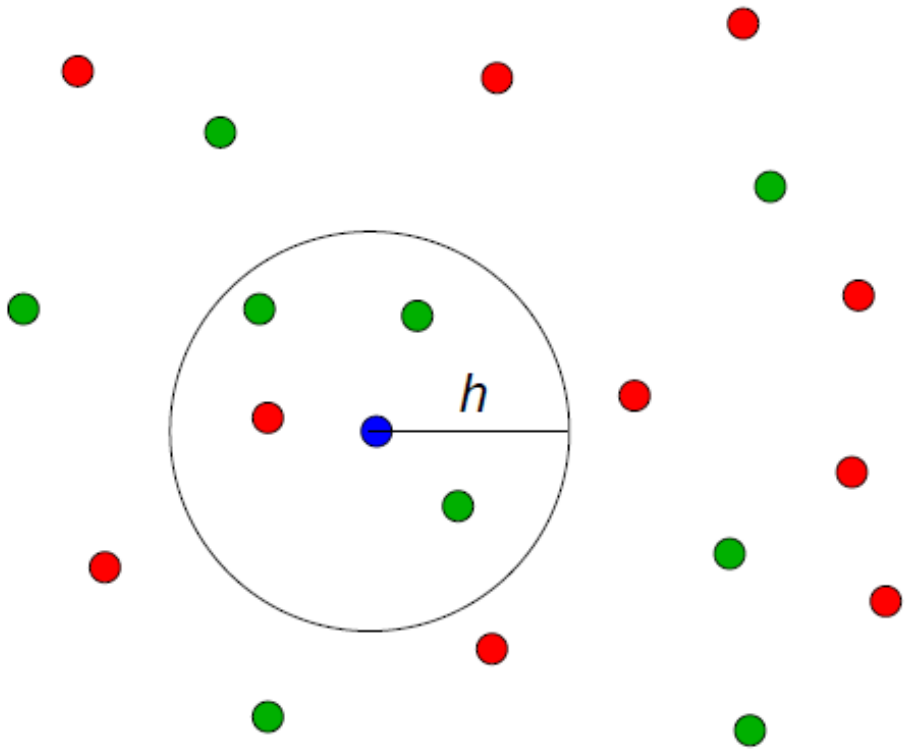



Figure: General behaviour of the Approximation risk and Estimation risk with respect to the complexity of the model

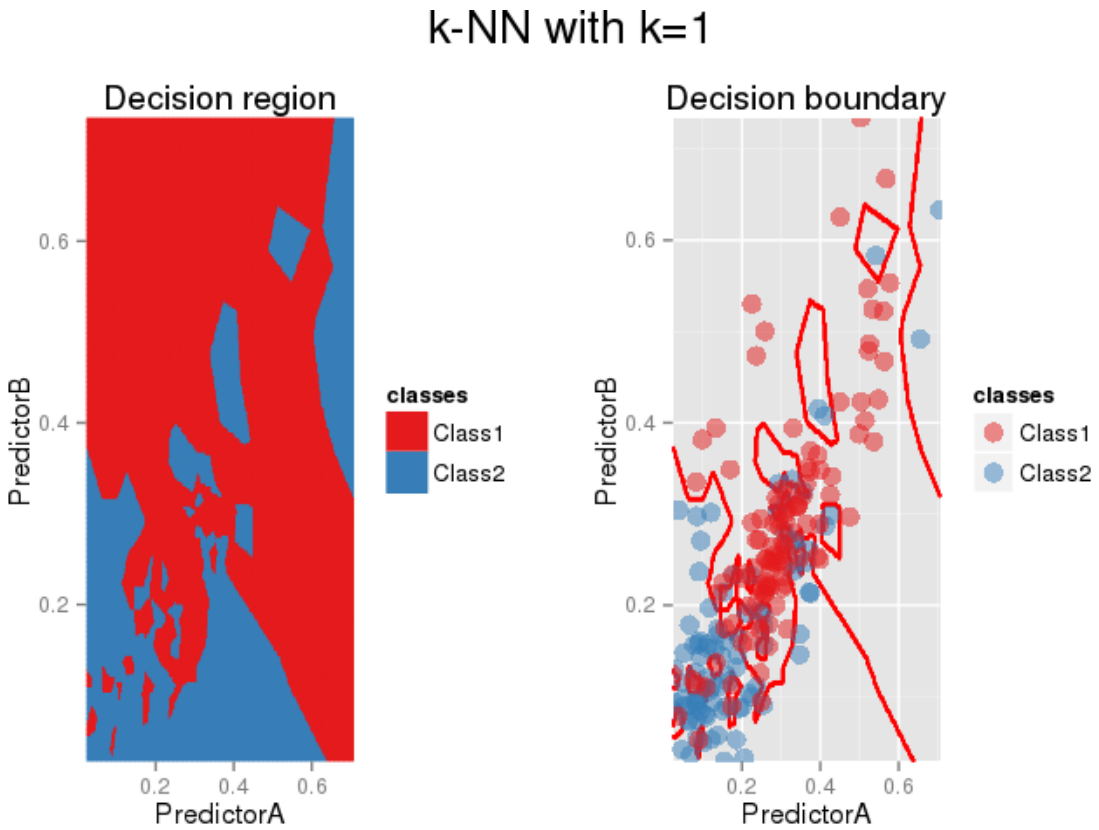
# Additional Example: $k$- nearest neighbor (with $k = 3$)



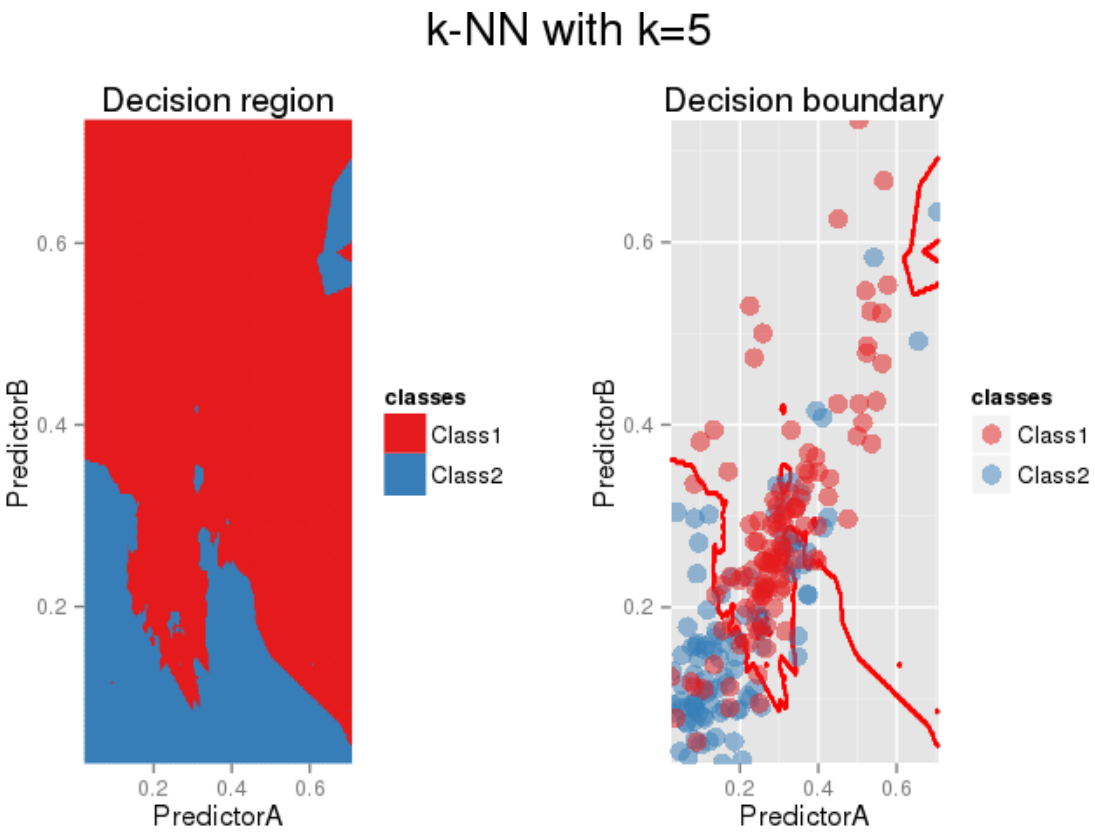
# Additional Example: $k$- nearest neighbor (with $k = 4$)



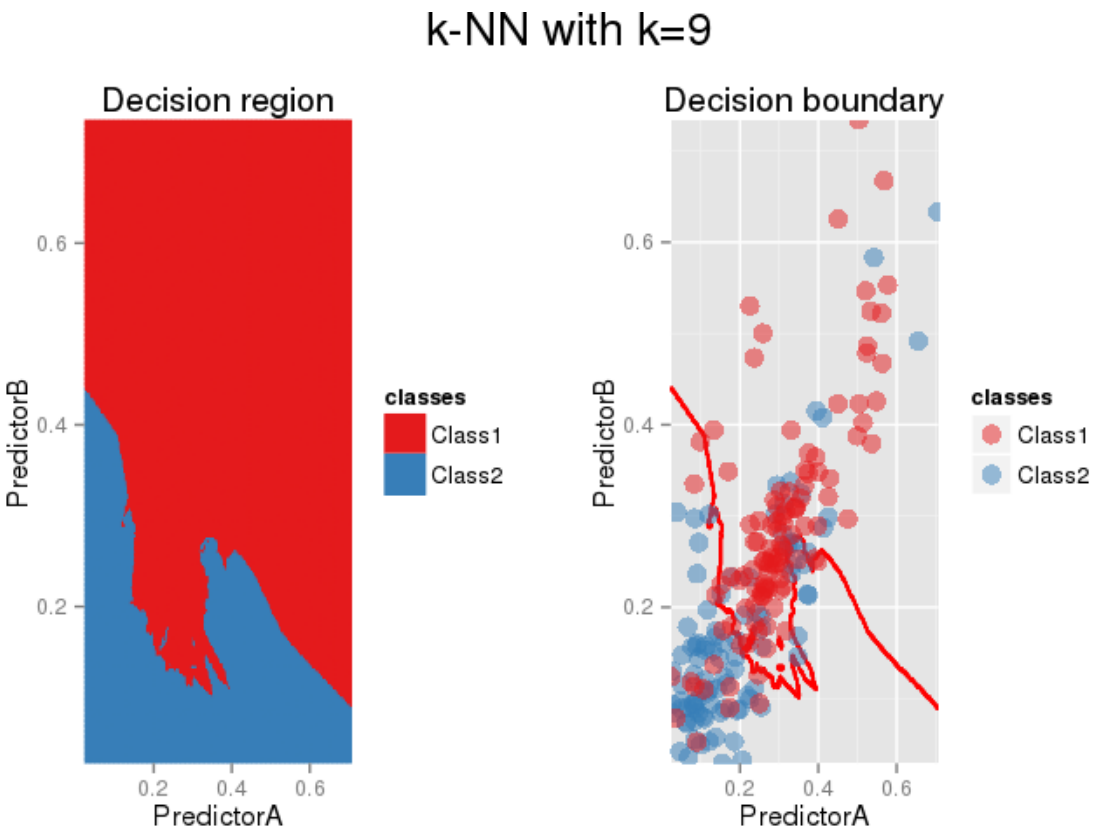
# Example: KNN I



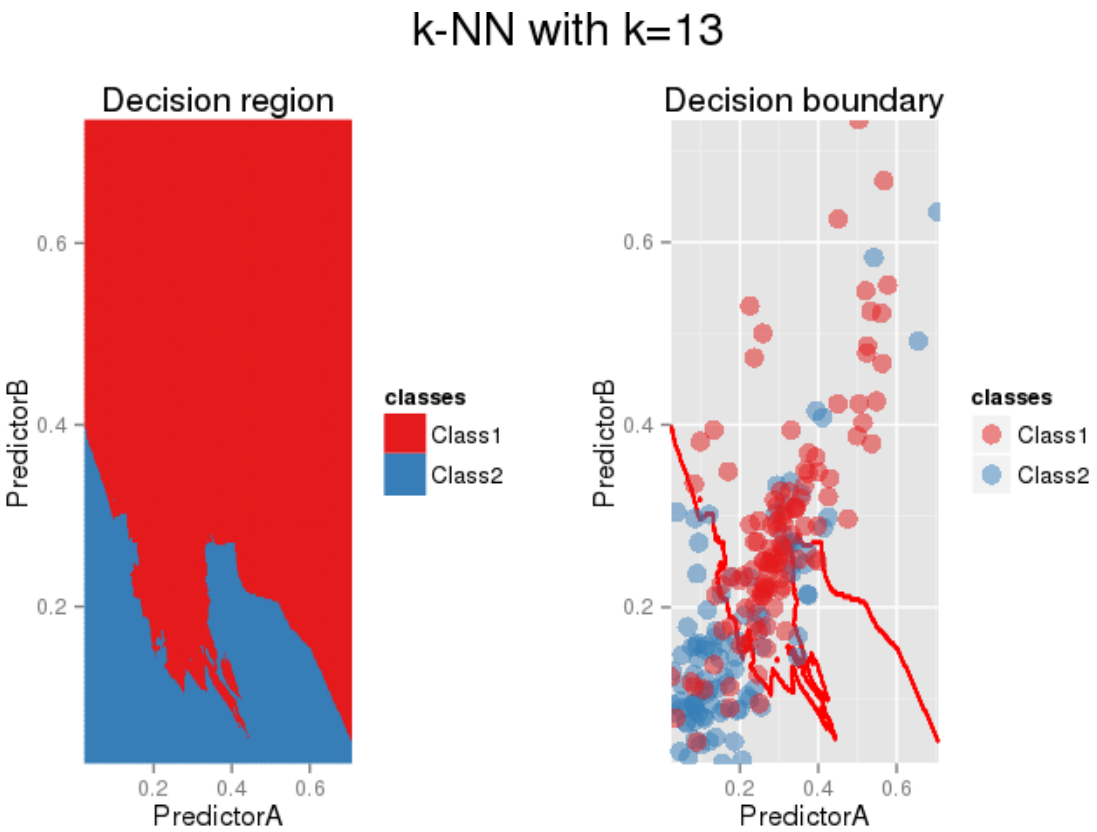
# Example: KNN II



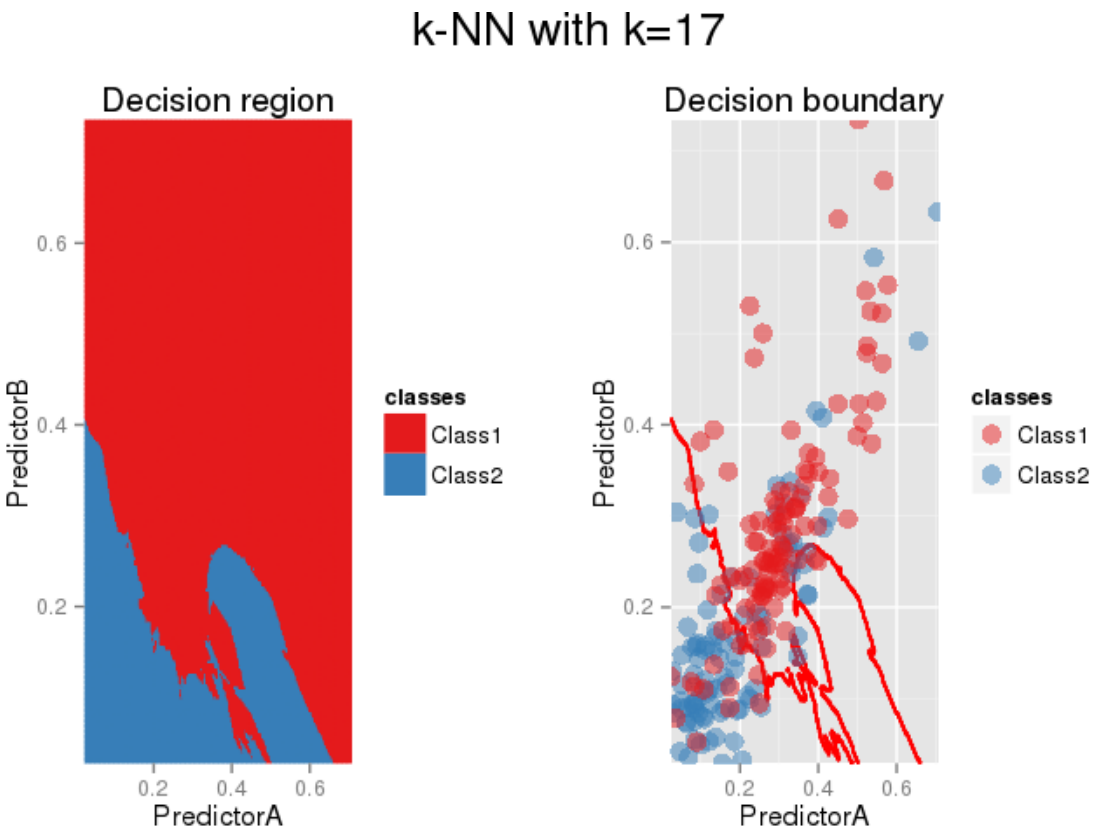
# Example: KNN III



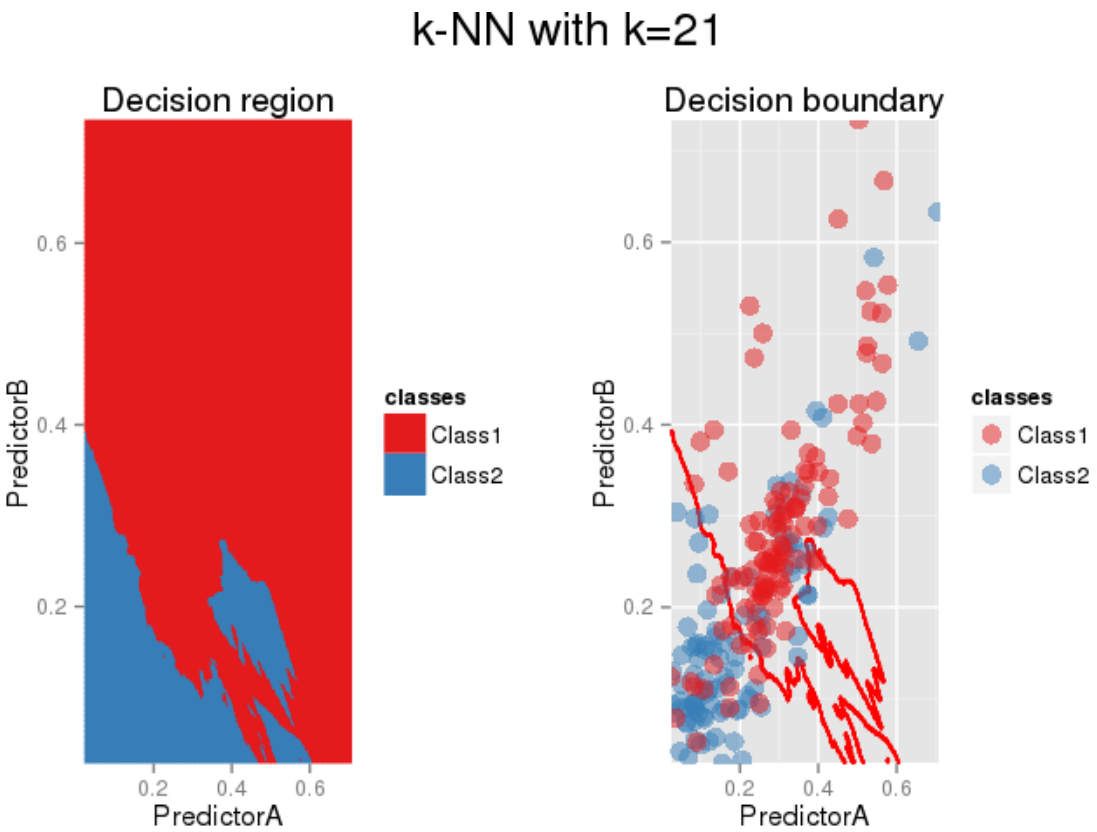
# Example: KNN IV



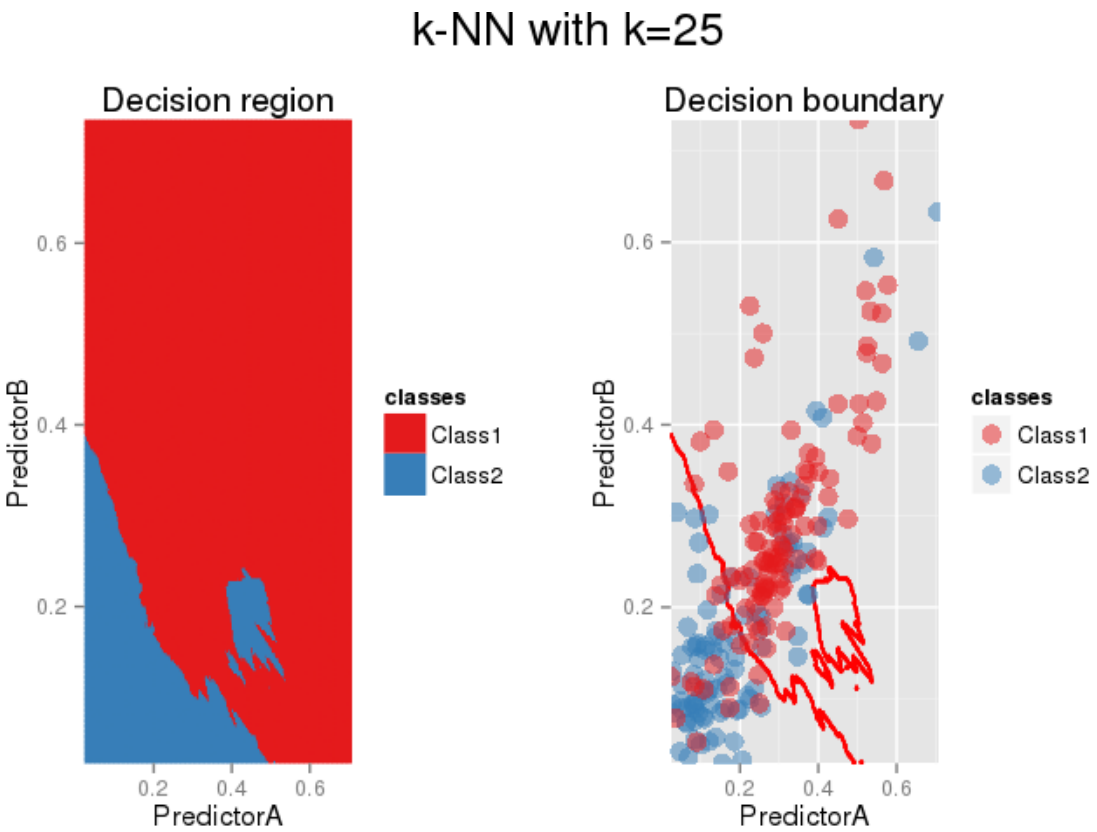
# Example: KNN V



# Example: KNN VI



# Example: KNN VII



# Example: KNN VIII