Machine Learning for biology

V. Monbet

UFR de Mathématiques
Université de Rennes 1

Outline

Outline

Supervised learning

- Problem: prediction of a variable $Y \in \mathcal{Y}$ given inputs $\mathbf{X} = (X_1, \cdots, X_p) \in \mathcal{X}$
- In practice, n observations of $Y \in \mathcal{Y}$ and $\mathbf{X} \in \mathcal{X}$ are available to learn the prediction "model". $\mathcal{S}_n = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$

Response (or output) Inputs

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{i1} & \cdots & x_{ip} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- The response Y can be quantitative (ex: Ozone concentration, El Niño index, temperature, ...) or qualitative (ex: yes/no, wet/dry, ...).
- Problem: find a mapping f such that

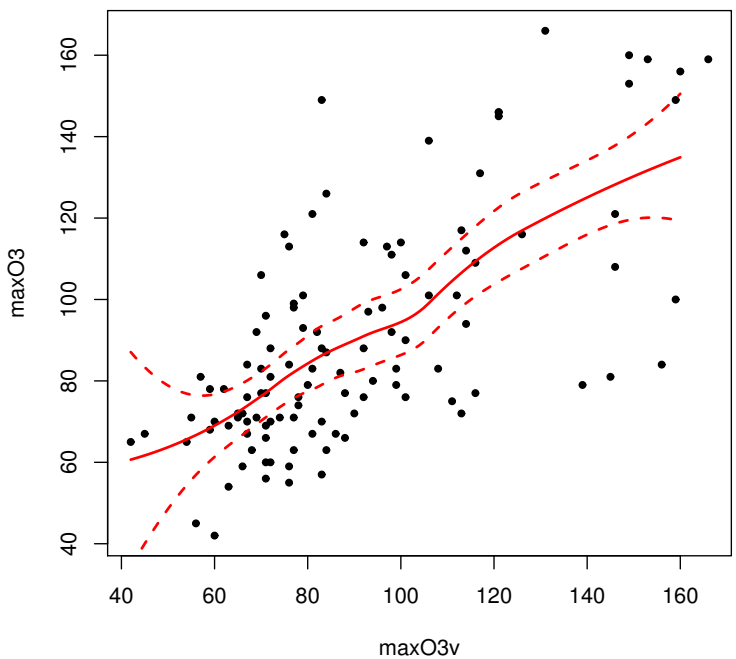
$$Y = f(\mathbf{X}) + \epsilon$$

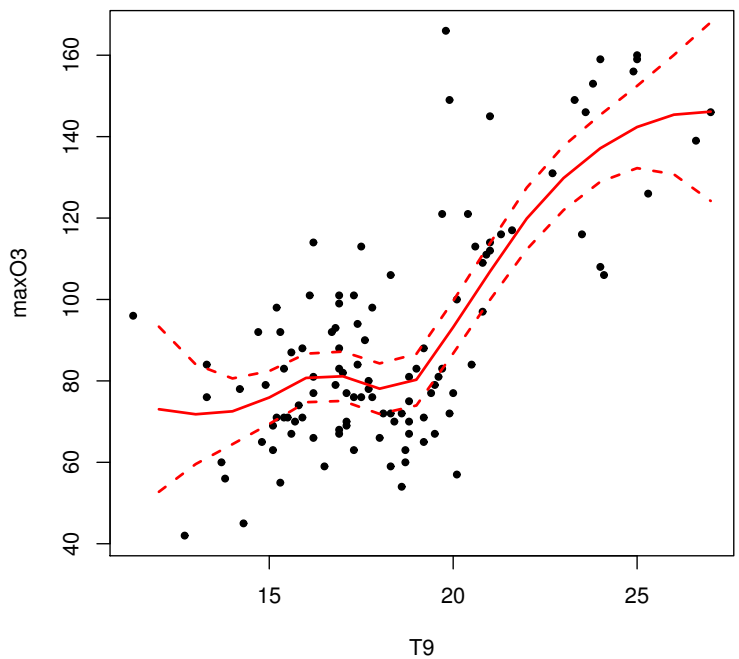
where *epsilon* denotes an error.

Supervised learning, example (regression)

Regression = prediction of a continuous variable.

- Example: predict maximum Ozone (O3) concentration on day D given maxO3 concentration and meteorological variables of the day before (temperature, nebulosity, West-East wind intensity, wind direction, rain yes/no) at times (9:00, 12:00, 15:00).
- Example: prediction of maxO3 given maxO3 the day before (left) and the temperature at 9:00 (right).
  Black dots correspond to observations, red solid lines to prediction and red dotted lines to a 95% confidence interval for the prediction.





Supervised learning, example (classification)

Classification = prediction of a categorical variable.

▶ Approximate the mapping : $f($  $) =$ Tomate

$$f : \mathbb{R}^{128 \times 128} \rightarrow \{\text{Apple, Pear, Tomato, Cow, Dog, Horse}\}$$

QUIZZ - Question 1

Suppose you want to develop a supervised machine learning model to predict whether a given email is "spam" or "not spam." Which of the following statements are true?

1. We'll use unlabeled examples to train the model.
2. The labels applied to some examples might be unreliable.
3. Emails not marked as "spam" or "not spam" are unlabeled examples.
4. Words in the subject header will make good labels.

QUIZZ - Response 1

Suppose you want to develop a supervised machine learning model to predict whether a given email is "spam" or "not spam." Which of the following statements are true?

1. We'll use unlabeled examples to train the model.
2. The labels applied to some examples might be unreliable.
   Definitely. It's important to check how reliable your data is. The labels for this dataset probably come from email users who mark particular email messages as spam. Since most users do not mark every suspicious email message as spam, we may have trouble knowing whether an email is spam. Furthermore, spammers could intentionally poison our model by providing faulty labels.
3. Emails not marked as "spam" or "not spam" are unlabeled examples.
   Because our label consists of the values "spam" and "not spam", any email not yet marked as spam or not spam is an unlabeled example.
4. Words in the subject header will make good labels.

QUIZZ - Question 2

Suppose an online shoe store wants to create a supervised ML model that will provide personalized shoe recommendations to users. That is, the model will recommend certain pairs of shoes to Marty and different pairs of shoes to Janet. Which of the following statements are true?

1. The shoes that a user adores is a useful label.
2. Shoe beauty is a useful feature.
3. Shoe size is a useful feature.
4. User clicks on a shoe's description is a useful label.

QUIZZ - Response 2

Suppose an online shoe store wants to create a supervised ML model that will provide personalized shoe recommendations to users. That is, the model will recommend certain pairs of shoes to Marty and different pairs of shoes to Janet. Which of the following statements are true?

1. The shoes that a user adores is a useful label.
2. Shoe beauty is a useful feature.
3. Shoe size is a useful feature.
   Shoe size is a quantifiable signal that likely has a strong impact on whether the user will like the recommended shoes. For example, if Marty wears size 9, the model shouldn't recommend size 7 shoes.
4. User clicks on a shoe's description is a useful label.
   Users probably only want to read more about those shoes that they like. User clicks is, therefore, an observable, quantifiable metric that could serve as a good training label.

Supervised learning

- Supervised learning methods can be mainly gathered in two groups: purely data driven approaches and methods based on a parametric model.

- **Data driven approaches** (non parametric)
  • Analogs (or k-nearest neighbors) consists in learning from similar (or dissimilar) observations.
  • Regression trees allow to split the space $\mathcal{X}$ of $\mathbf{x}$ into small regions with constant y.

- **Modeling approaches** (parametric)
  • Linear regression
  • Artificial Neural Network and Deep learning
  • Support Vector Machines

- **Aggregation models**: Some of these methods or models can be combined to take advantage of the best features of each of them.
  • Bagging and random forest
  • Boosting, Gradient Boosting

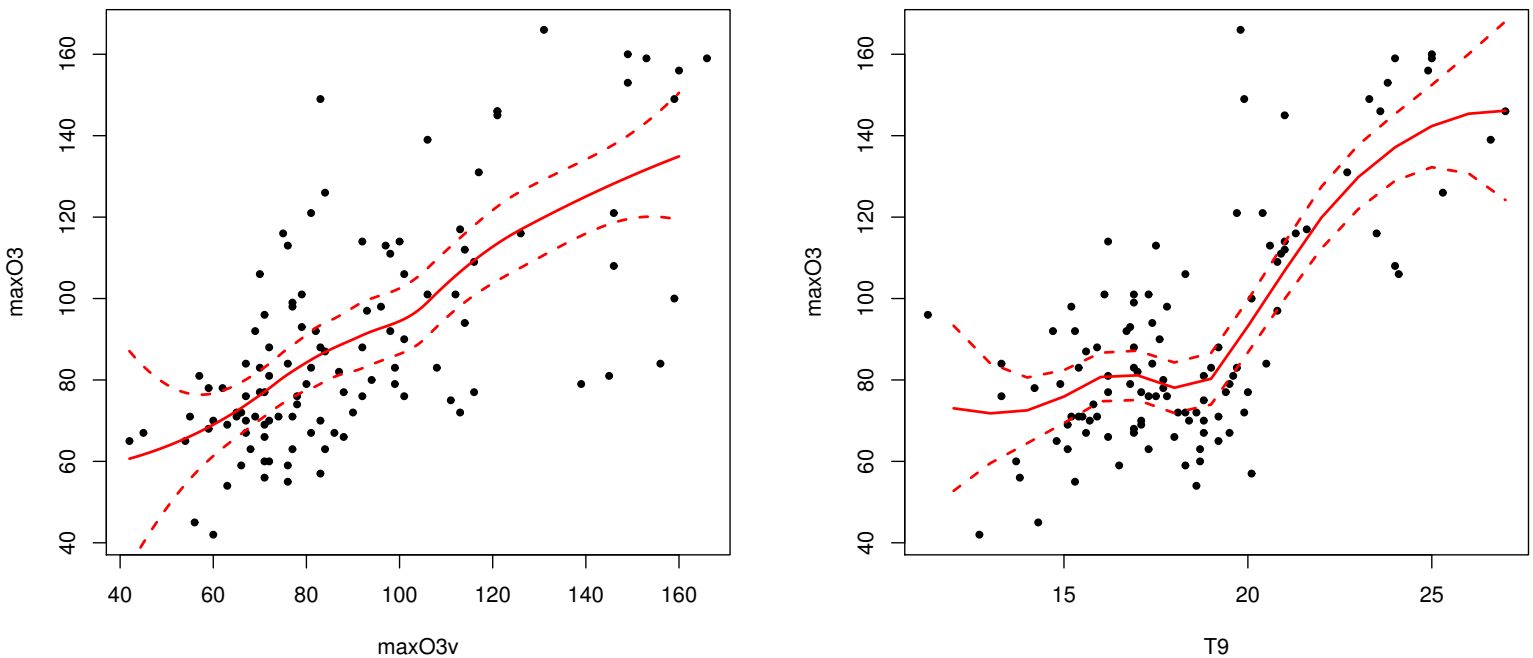
Outline

Outline

Supervised learning, example (regression)

Regression = prediction of a continuous variable.

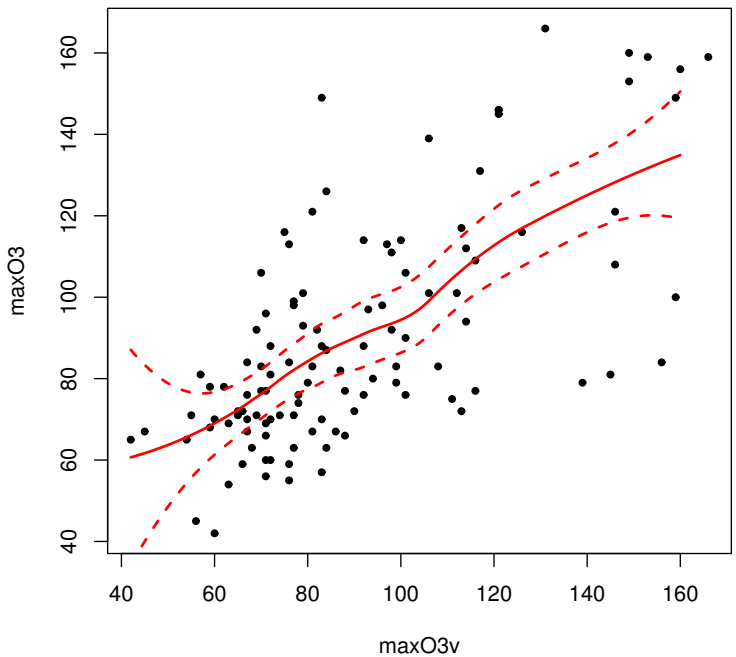
- Example: predict maximum Ozone (O3) concentration on day D given maxO3 concentration and meteorological variables of the day before (temperature, nebulosity, West-East wind intensity, wind direction, rain yes/no) at times (9:00, 12:00, 15:00). Black dots correspond to observations, red solid lines to prediction and red dotted lines to a 95% confidence interval for the prediction.

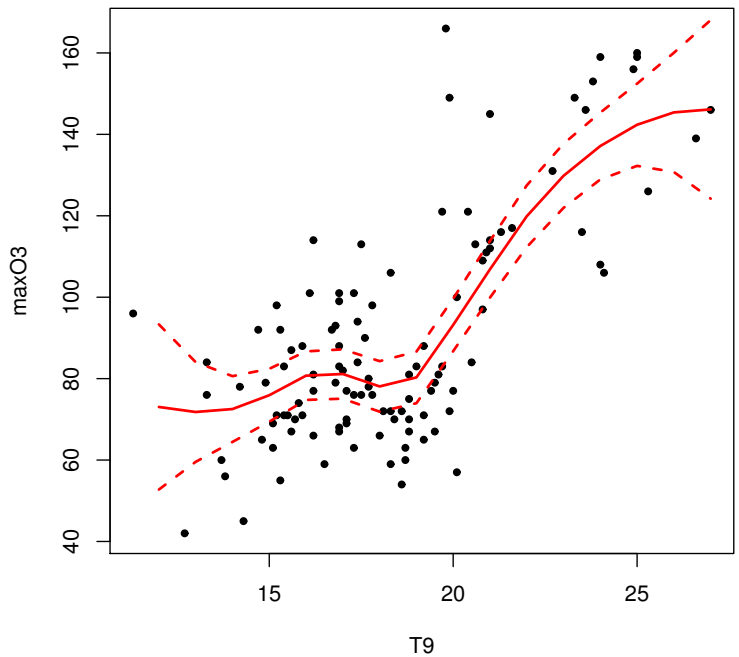


- Remark : the prediction (red line) materializes a local mean. This local mean will be denoted $E(Y|\mathbf{X} = \mathbf{x})$ in the sequel.

Supervised learning, example (regression)

Regression = prediction of a continuous variable.
Black dots correspond to observations, red solid lines to prediction and red dotted lines to a 95% confidence interval for the prediction.





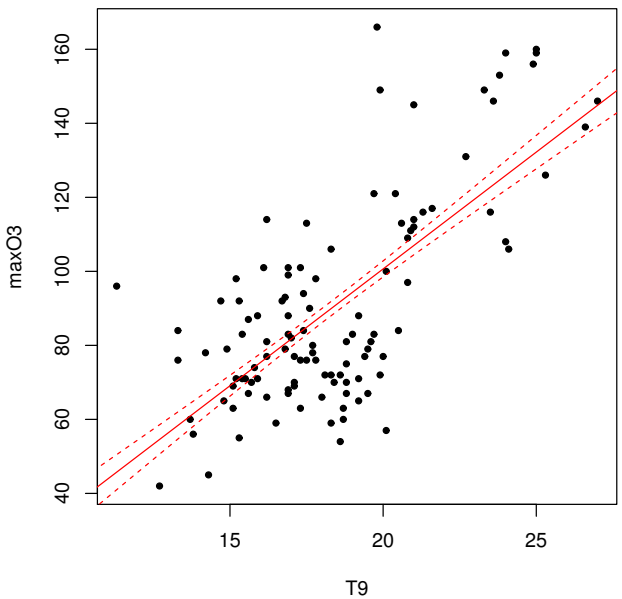
Remark : the prediction (red line) materializes a local mean.
This local mean will be denoted $E(Y|\mathbf{X} = \mathbf{x})$ in the sequel.

Standard machine learning methods only focus on the local mean (and give no information about the incertainty around the mean.

Linear regression models

- A linear regression model assumes that the regression function $f(x) = E(Y|\mathbf{X} = \mathbf{x})$ is linear in the inputs $X_1, \cdots, X_p$.

It implies that f is written as follows.



$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \mathbf{X}\beta$$

- They are simple and often provide an adequate and interpretable description of how the inputs affect the output.
- For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data.
- Finally, linear methods can be applied to transformations of the inputs and this considerably expands their scope.

Linear regression models

- We have an input vector $\mathbf{X} \in \mathcal{X}$, and want to predict a real-valued output $Y \in \mathcal{Y} \subseteq \mathbb{R}$.
- The linear regression model has the form

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \mathbf{X}\beta \tag{1}$$

why $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \mathbf{X}\beta$?

- Here the β_j's are unknown parameters or coefficients, and the variables X_j can come from different sources (numerical, categorical).
- The most popular estimation method is least squares in which we pick the coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ to minimize the residual sum of squares (sum of squared residuals =mean square error (MSE))

$$MSE(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

given a learning dataset $\mathcal{S} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$

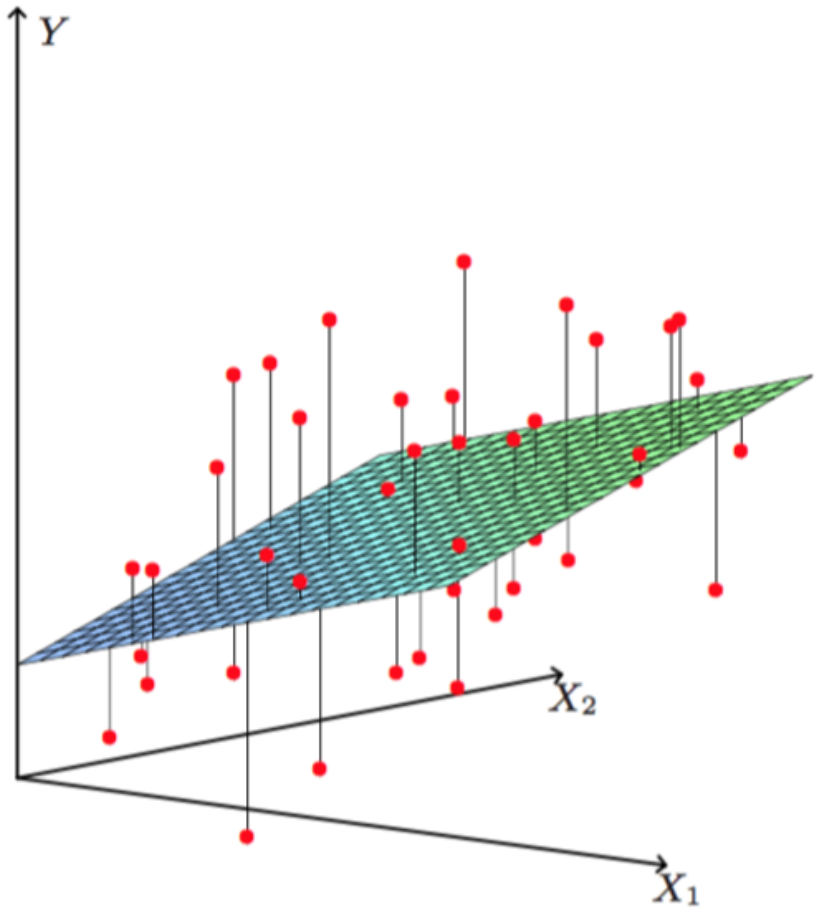


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Figure from Hastie's book.

Minimization of RSS and prediction

- One can rewrite *RSS* in a matricial form with $\mathbf{x} = [\mathbf{1}, \mathbf{x}]$

$$RSS(\beta) = (y - \mathbf{x}\beta)^T (y - \mathbf{x}\beta).$$

- This is a quadratic function in the p parameters vector β. Differentiating with respect to β Differentiation of $(y - \mathbf{x}\beta)^T (y - \mathbf{x}\beta)$? and setting the derivatives to zero, one obtains estimators

$$\widehat{\beta} = \left(\mathbf{x}^T \mathbf{x}\right)^{-1} \mathbf{x}^T y$$

This is an exact formula easy to compute (under some conditions).

- For prediction, β is substitue by $\widehat{\beta}$ in Eq. (1)

$$\hat{y} = \mathbf{x}\widehat{\beta} = \mathbf{x}\left(\mathbf{x}^T \mathbf{x}\right)^{-1} \mathbf{x}^T y$$

- Model predictive power is often measured by the percentage of explained variance

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\left(\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{y})^2\right)^{1/2}}$$

$R^2 = 1$ means that $\hat{y} = y$ and $R^2 = 0$ means that $\widehat{\beta} = (\bar{y}, 0, \cdots, 0)$

Important remarks/summary

- As many machine learning methods, the linear model is fitted by minimizing the mean square error. Note that it is a global criteria ie all the observations are involved in the criteria.
- The specificity of linear model is the shape constraint imposed to the regression fonction.
- Estimators are obtained by a close form expression

$$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y$$

$\rightarrow$ **no numerical optimization algorithm has to be run**

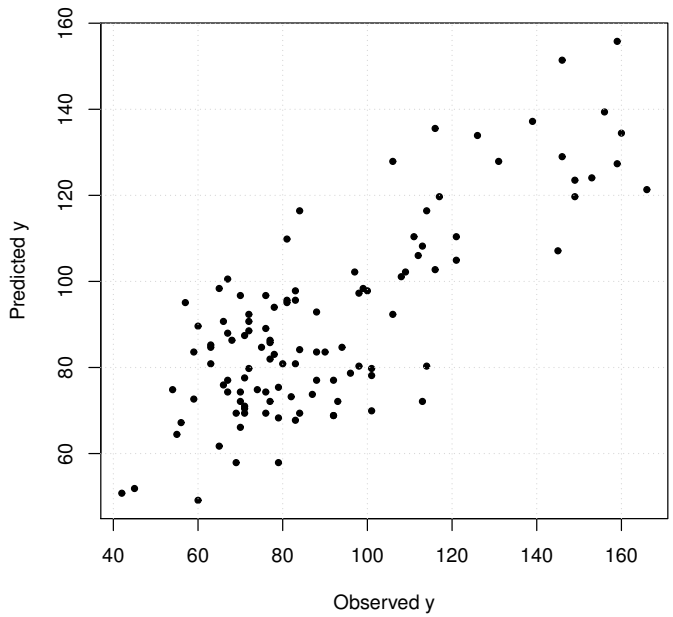
- Linear models are elementary blocks for other machine learning models such as neural networks, deep learning, machine support vectors.

Validation of a regression model

Example: Ozone

To validate a regression mode the first step is to plot the prediction with respect to the observation. It is a qualitive validation.

- Model: $max03 = \beta_0 + \beta_1 T12$
  R^2 = 0.61, R^2_{aj} = 0.48



Quantitative measures are usefull too because they provide more objective criteria.

Validation of a regression model, quantitative measures

- The quality of the model is usually measured by the square root of the mean/average of the square of all of the errors (root mean square error).

$$RMSE = \sqrt{\sum_{i \in \mathcal{I}} (\hat{y}_i - y_i)^2}$$

  where $\mathcal{I}$ denotes a set of validation individuals (see below).
- But, RMSE is very sensible to few extreme errors.
- Some other loss functions may be defined, as Mean Absolute Error

$$MAE = \sum_{i \in \mathcal{I}} |\hat{y}_i - y_i|$$

Outline

Cross-validation

- **Cross-validation** is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.
  The quality of the model is usually measured by the square root of the mean/average of the square of all of the error (root mean square error).

$$RMSE = \sqrt{\sum_{i \in \mathcal{I}} (\hat{y}_i - y_i)^2}$$

- It is also used to calibrate algorithm parameters (e.g. number of neighbors k, variable selection).
- The goal of cross validation is to define a dataset to "test" the model in order to limit problems like overfitting.
- In practice a partition of the dataset into 2 subsets is generated randomly. The model is calibrated on the first part and tested/validated on the second part. And this is repeated.

Cross-validation, algorithms

K-fold CV

index = sample n **indices in** $1, 2, \cdots, n$ **without replacement**

$n_K = n/K$

for $k = 1$ **to** K**,**

$train$ **= index**$[k * n_K + (1 : n_K)]$

$test = \{1, \cdots, n\} - \{train\}$

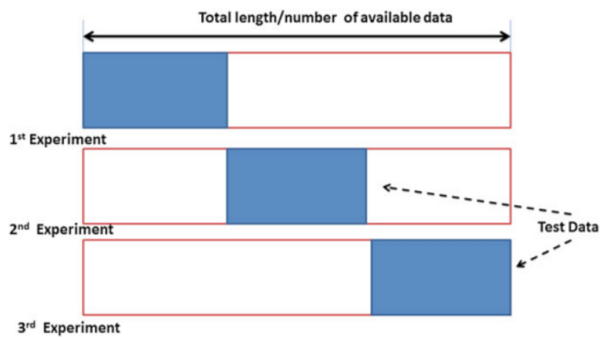
Calibrate the model M_k **on** $train$

Predict $\hat{Y}_i = M_k(\mathbf{x}_i)$**,** $\forall i \in test$

$MSE(k) = \frac{1}{card(test)} \sum_{i \in test} (\hat{Y}_i - y_i)^2$

end for

$RMSE = \sqrt{\frac{1}{K} \sum_{k=1}^{K} MSE(k)}$



Monte Carlo CV

$n_{train} = 2/3 * n$

$B = 100$

for $b = 1$ **to** B**,**

$train$ **= sample** n_{train} **indices in** $1, 2, \cdots, n$ **without replacement**

$test = \{1, \cdots, n\} - \{train\}$

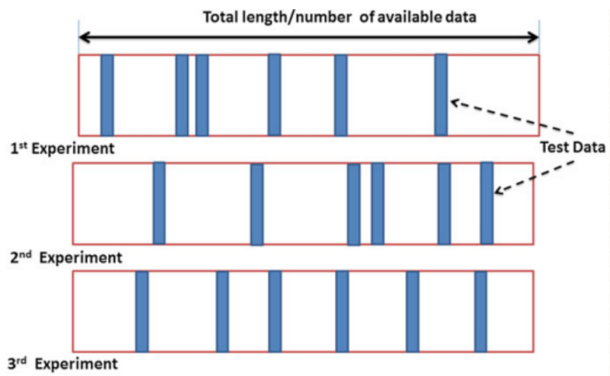
Calibrate the model M_k **on** $train$

Predict $\hat{Y}_i = M_k(\mathbf{x}_i)$**,** $\forall i \in test$

$MSE(b) = \frac{1}{card(test)} \sum_{i \in test} (\hat{Y}_i - y_i)^2$

end for

$RMSE = \sqrt{\frac{1}{B} \sum_{b=1}^{B} MSE(b)}$



k-fold CV vs. Monte Carlo CV

- Under k-fold cross validation, each point gets tested exactly once, which seems fair. However, cross-validation only explores a few of the possible ways that your data could have been partitioned.
- Monte Carlo lets you explore somewhat more possible partitions, though you're unlikely to get all of them.
- Averaging the results of a *k*-fold cross validation run gets you a (nearly) unbiased estimate of the algorithm's performance, but with high variance (as you'd expect from having only 5 or 10 data points).
- Since you can, in principle, run it for as long as you want/can afford, Monte Carlo cross validation can give you a less variable, but more biased estimate.
- Best choice: Monte Carlo cross validation

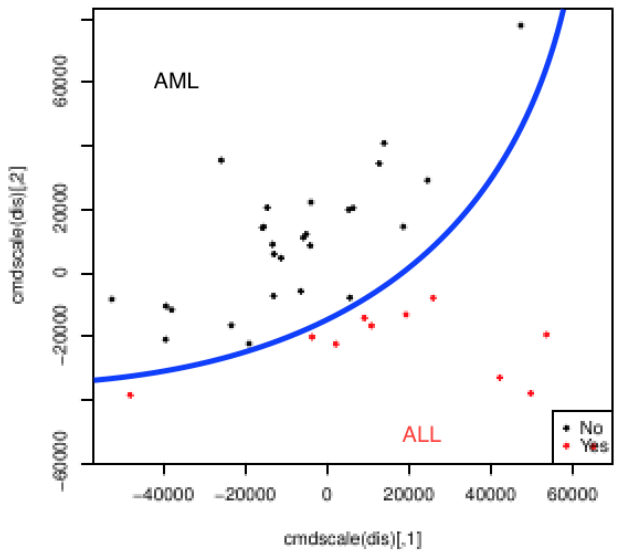
Outline

Extensions of linear model for categorical variable

- When Y is a categorical variable the prediction problem is equivalent to search for boundaries between classes.
- There are several different (generalized) linear [a] approaches to model the boundaries
- The most common methods are linear discriminant analysis or generalized linear models (ex: logistic regression).

[a] linear means that the decision rule can be expressed as $g(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$

Here, the problem is to predict a leucemia type from microarray data. The data have been projeted on a 2 dimensions space by MDS, after a preselection of the most informative genes.



Logistic regression

- Logistic regression is a direct and straighforward extension of linear regression.
- $Y \in \{0, 1\}$ is binary.
- So that $P(Y = 1|\mathbf{X} = \mathbf{x}) \in [0, 1]$.
- A link function $g : \mathbf{R} \mapsto [0, 1]$ is introduced

$$P(Y = 1) = g\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_j\right)$$

In the logistic model,

$$g(x) = \exp(x)/(1 + \exp(x))$$

- Parameter β is estimated by maximum likelihood.

Validation of classification problems

Confusion Matrix

Let's make the following definitions:
- "Pollution" is a positive class.
- "No pollution" is a negative class.
We can summarize our "pollution-prediction" model using a 2×2 confusion matrix that depicts all four possible outcomes:

True Positive (TP): - Reality : the river is polluted - Model prediction: pollution	False Positive (FP): - Reality : the river is not polluted - Model prediction: pollution
False Negative (FN): - Reality : the river is polluted - Model prediction: no pollution	True Negative (TN): - Reality : the river is not polluted - Model prediction: no pollution

- A **true positive** is an outcome where the model correctly predicts the positive class. Similarly, a **true negative** is an outcome where the model correctly predicts the negative class.
- A **false positive** is an outcome where the model incorrectly predicts the positive class. And a **false negative** is an outcome where the model incorrectly predicts the negative class.

Accuracy

Informally, **accuracy** is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$\text{accuracy} = \frac{\text{number of correct prediction}}{\text{total of number prediction}}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

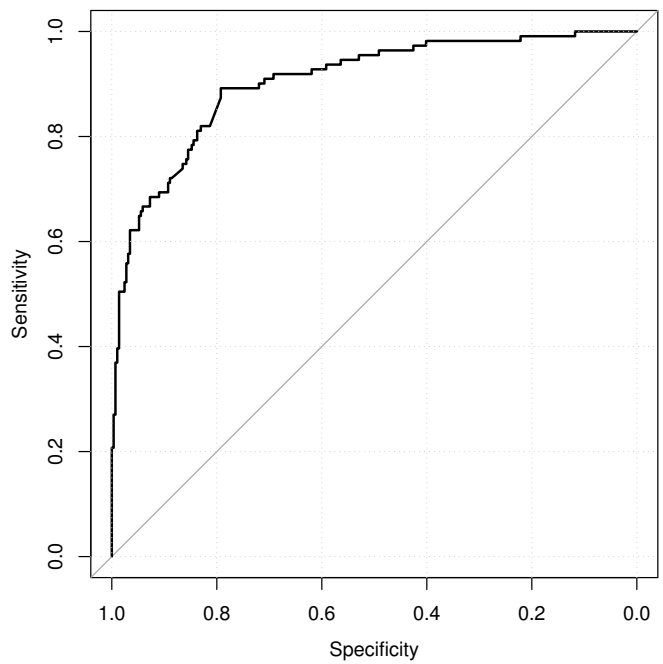
Accuracy alone doesn't tell the full story when you're working with a class-imbalanced data set, like this one, where there is a significant disparity between the number of positive and negative labels.

ROC curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate (or sensitivity) $TPR = \frac{TP}{TP+FN}$
- False Positive Rate (or specificity) $FPR = \frac{FP}{FP+TN}$.

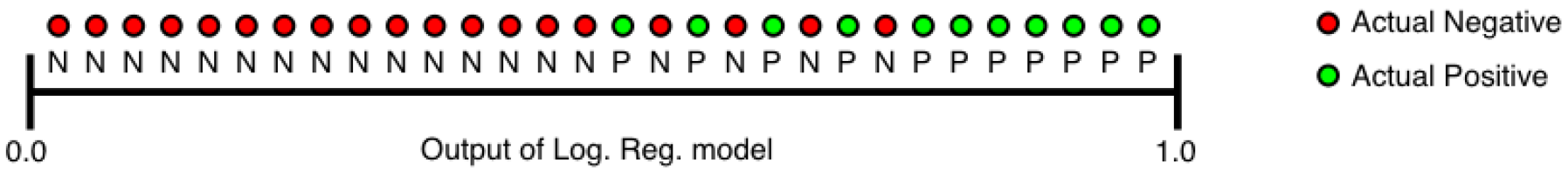
An ROC curve plots TPR vs. FPR at different classification thresholds.



Area under the ROC curve (AUC)

AUC is the area under the ROC curve.

- $0 \leq AUC \leq 1$. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.
- The largest the best.
- AUC is scale invariant. It measures how well predictions ($P(\widehat{Y = 1})$) are ranked, rather than their absolute values.
- AUC is threshold invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.



QUIZZ - Question

In which of the following scenarios would a high accuracy value suggest that the ML model is doing a good job?

1. A deadly, but curable, medical condition afflicts .01% of the population. An ML model uses symptoms as features and predicts this affliction with an accuracy of 99.99%.
2. In the game of roulette, a ball is dropped on a spinning wheel and eventually lands in one of 38 slots. Using visual features (the spin of the ball, the position of the wheel when the ball was dropped, the height of the ball over the wheel), an ML model can predict the slot that the ball will land in with an accuracy of 4%.
3. An expensive robotic chicken crosses a very busy road a thousand times per day. An ML model evaluates traffic patterns and predicts when this chicken can safely cross the street with an accuracy of 99.99%.

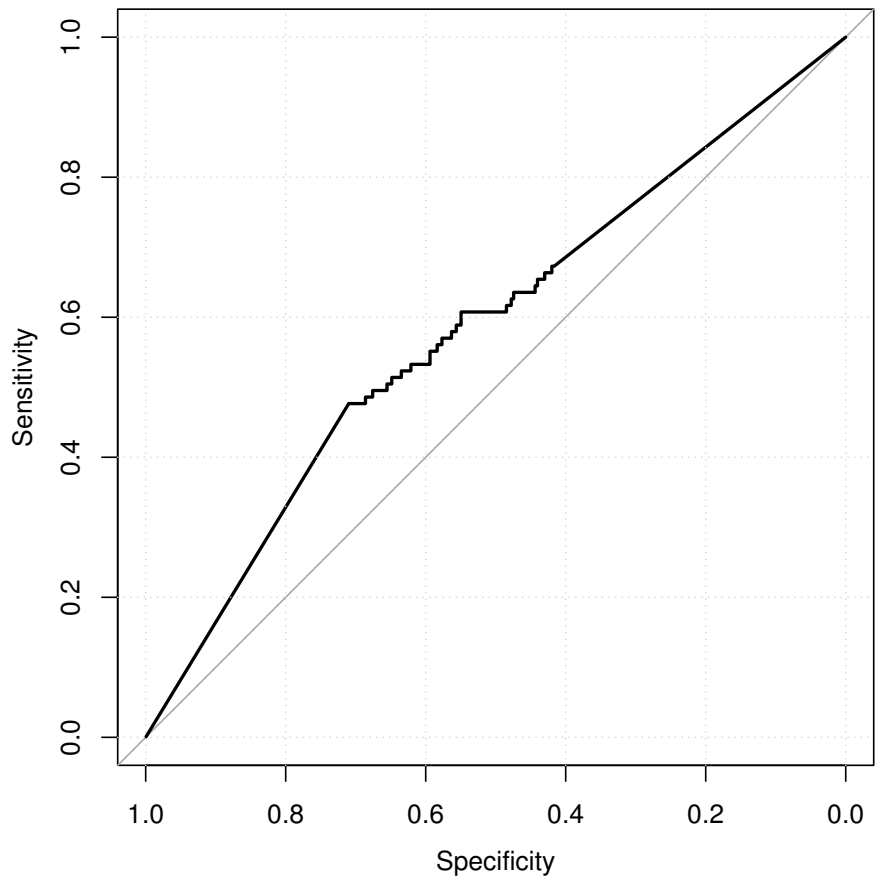
QUIZZ - Response

In which of the following scenarios would a high accuracy value suggest that the ML model is doing a good job?

1. A deadly, but curable, medical condition afflicts .01% of the population. An ML model uses symptoms as features and predicts this affliction with an accuracy of 99.99%.
2. In the game of roulette, a ball is dropped on a spinning wheel and eventually lands in one of 38 slots. Using visual features (the spin of the ball, the position of the wheel when the ball was dropped, the height of the ball over the wheel), an ML model can predict the slot that the ball will land in with an accuracy of 4%.
   This ML model is making predictions far better than chance; a random guess would be correct 1/38 of the time-yielding an accuracy of 2.6%. Although the model's accuracy is "only" 4%, the benefits of success far outweigh the disadvantages of failure.
3. An expensive robotic chicken crosses a very busy road a thousand times per day. An ML model evaluates traffic patterns and predicts when this chicken can safely cross the street with an accuracy of 99.99%.

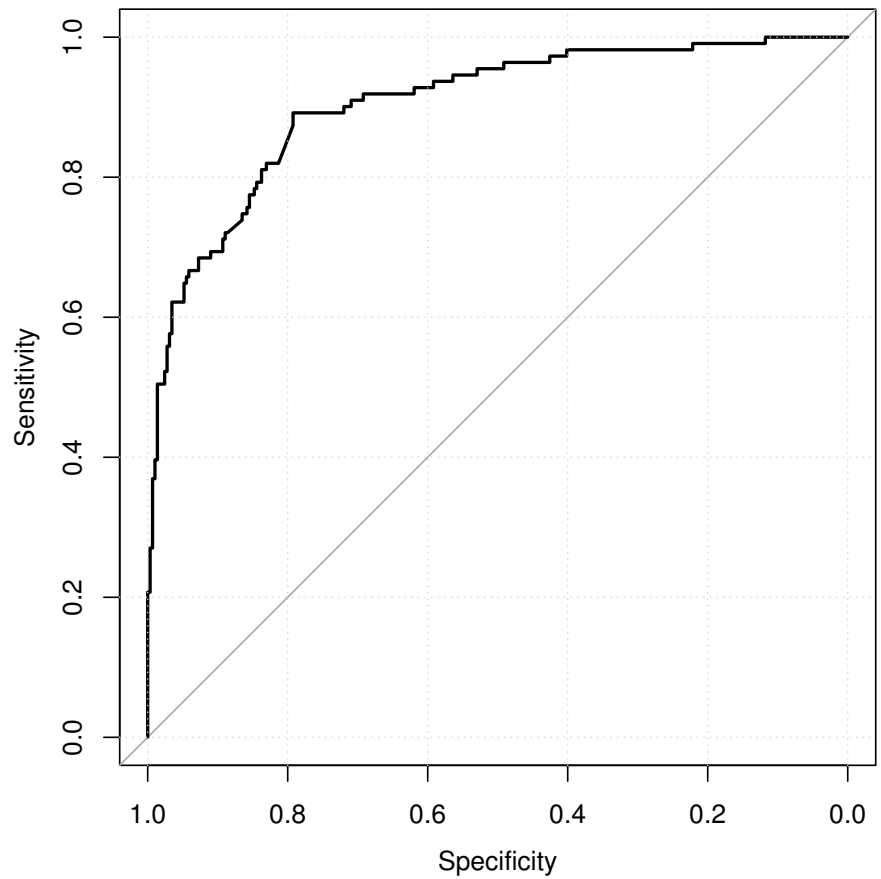
Logistic regression, example

- Leukemia gene expression
- Model based on
- the genes with highest absolute correlation with class AML/ALL ($|\rho| > .5$)
- n.train = 34, AUC = 0.58
- When the dimension is high, overfitting occurs in logistic regression: in the training step a perfect solution is found which is not robuts.



Logistic regression, LASSO

- Leukemia gene expression
- A LASSO penalty can be added at the estimation step to reduce overfitting.
- Model based on the genes with highest absolute correlation with class AML/ALL ($|\rho| > .5$)
- n.train = 34, AUC = 0.90



Recap

- **Linear model** is used to predict a quantitative variable Y by a linear combination of the X variables.
  X variables can be quantitative or qualitative.
- The estimators are easy to computed. When $p < n$, it usually gives a good first approximation of the relationship between X and Y.

- **If Y is qualitative**, generalized linear model are considered such as logistic regression.
- Linear discriminant analysis is a reduction dimension like technique which leads to linear frontiers between classes defined by Y.
  Quadratic discriminant analysis is an extension for quadratic frontiers.

- **Validation** is based on RMS/MAE if Y is quantitative and classification error/confusion matrix/ROC if Y is qualitative.
- Whatever the case, cross validation has to be used for fair model validation.

Bayesian decision rule

- In classification problems, Y takes its values in $\{1, \cdots, K\}$ with probabilities $\pi_1, \cdots, \pi_K$.
- $\pi_\ell = P(Y = \ell)$ correspond to the *a priori probabilities* of classes (ex: patient/control).
- Now let us suppose that the predictors $\mathbf{X}$ have probability density functions

$$f_\ell(\mathbf{x}) = P[\mathbf{X}|Y = \ell]$$

in each class ℓ.

- The decision rule of LDA is based on the *posterior probabilities*

$$P(Y = \ell|\mathbf{X} = \mathbf{x})$$

- **Bayes rule**
A sample corresponding to observation $\mathbf{x}$ is classified in class ℓ if

$$P(Y = \ell|X = x) \geq P(Y = k|\mathbf{X} = \mathbf{x}) \text{ for all } k \in \{1, \cdots, K\}$$

- The Bayes rule is optimal for the classification error.
- A simple application of Bayes theorem Bayes formula gives us

$$P(Y = \ell|\mathbf{X} = \mathbf{x}) = \frac{\pi_\ell f_\ell(\mathbf{x})}{\sum_{k=1}^K \pi_k f_k(\mathbf{x})}$$

Linear discriminant analysis

- The linear discriminant analysis (LDA) is obtained when the densities $f_\ell(.)$ are Gaussian with the same covariance matrice Σ in all classes.

$$f_\ell(\mathbf{x}) = \frac{(2\pi)^{d/2}}{\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\ell)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ell)\right)$$

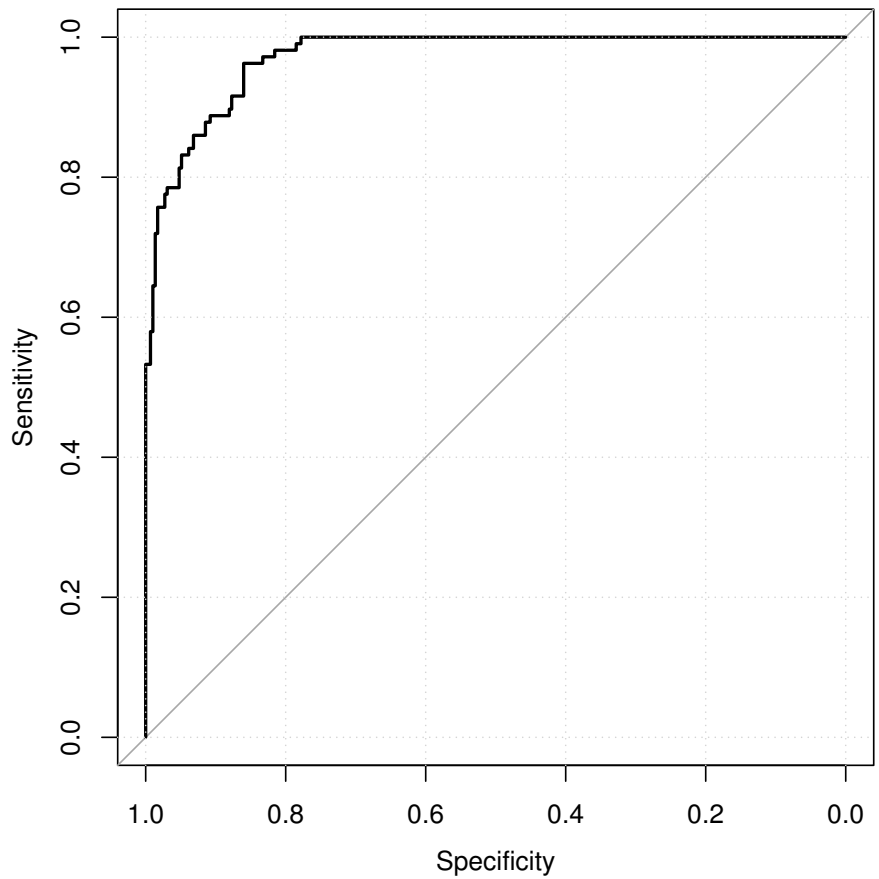
- The Bayes rule leads to a linear boundaries:

$$\begin{aligned}\log\frac{P(Y=k|\mathbf{X}=\mathbf{x})}{P(Y=\ell|\mathbf{X}=\mathbf{x})} &= \log\frac{f_k(\mathbf{x})}{f_\ell(\mathbf{x})} + \log\frac{\pi_k}{\pi_\ell} \\ &= \log\frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_\ell)^T\Sigma^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_\ell) \\ &\quad +\mathbf{x}^T\Sigma^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_\ell).\end{aligned}$$

  This equation is linear in $\mathbf{x}$.
- The quadratic discriminant analysis (QDA) is obtained when the densities $f_\ell(.)$ are Gaussian with the different covariance matrices Σ_ℓ in each classe.
- Naive Bayes models are variants of the discriminant analysis, and assume that each of the class densities are products of marginal Gaussian densities.

LDA, example

- Leukemia gene expression
- Model based on
- the genes with highest absolute correlation with class AML/ALL ($|\rho| > .5$)
- n.train = 34, AUC = 0.97

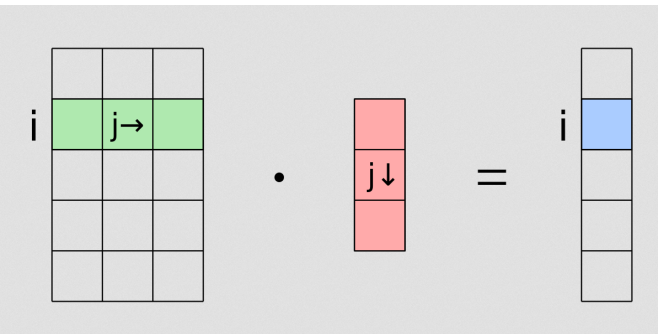


Produit matrice-vecteur

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \mathbf{X}\beta$$

Let $\mathbf{X}$ be defined as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$



then the line i of $\mathbf{X}\beta$ is

$$(\mathbf{X}\beta)_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

◂ back

Differentiation of the quadratic form

Firstly note that

$$(y - \mathbf{x}\beta)^T(y - \mathbf{x}\beta) = y^T y + y^T\mathbf{x}\beta + \beta^T\mathbf{x}^T y + \beta^T\mathbf{x}^T\mathbf{x}\beta$$

now because $\mathbf{x}\beta$ and y are vectors, one has

$$y^T\mathbf{x}\beta = \beta^T\mathbf{x}^T y$$

so that

$$(y - \mathbf{x}\beta)^T(y - \mathbf{x}\beta) = y^T y + 2y^T\mathbf{x}\beta + \beta^T\mathbf{x}^T\mathbf{x}\beta$$

1st left term does not depend on β, 2nd one depends linearly and the 3rd one is quadratic in β. And when, the derivative is calculated with respect to β

$$\frac{\partial(y - \mathbf{x}\beta)^T(y - \mathbf{x}\beta)}{\partial\beta} = 2y^T\mathbf{x} + 2\mathbf{x}^T\mathbf{x}\beta$$

◂ back

Bayes formula

The general Bayes formula is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

In the context of discriminant analysis,

$$P(Y = \ell | \mathbf{X} = \mathbf{x}) \quad = \quad \frac{P(Y = \ell \text{ and } \mathbf{X} = \mathbf{x})}{P(\mathbf{X} = \mathbf{x})}$$

and, by definition, $P(Y = \ell) = \pi_\ell$, $P(\mathbf{X} = \mathbf{x}|Y = \ell) = f_\ell(\mathbf{x})$
Now, by applying the Bayes formula $P(A \cap B) = P(A|B)P(B)$,

$$P(Y = \ell \text{ and } \mathbf{X} = \mathbf{x}) = P(\mathbf{X} = \mathbf{x}|Y = \ell)P(Y = \ell) = f_\ell(\mathbf{x})\pi_\ell$$

And, from the total probabilities formula,

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}) &= \sum_{k=1}^{K} P(\mathbf{X} = \mathbf{x}|Y = k)P(Y = k) \\ &= \frac{\pi_\ell f_\ell(\mathbf{x})}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x})} \end{aligned}$$

back

Outline