# 02450: Introduction to Machine Learning and Data Mining

Performance evaluation, Bayes, and Naive Bayes

Morten Mørup and Mikkel N. Schmidt

DTU Compute, Technical University of Denmark (DTU)

# Lecture Schedule

1. Introduction
   7 October: C1

Data: Feature extraction, and visualization

2. Data, feature extraction and PCA
   7 October: C2, C3
3. Measures of similarity, summary statistics and probabilities
   7 October: C4, C5
4. Probability densities and data Visualization
   7 October: C6, C7

Supervised learning: Classification and regression

5. Decision trees and linear regression
   8 October: C8, C9
6. Overfitting, cross-validation and Nearest Neighbor
   8 October: C10, C12
7. **Performance evaluation, Bayes, and Naive Bayes**
   **9 October: C11, C13**
8. Artificial Neural Networks and Bias/Variance
   9 October: C14, C15
9. AUC and ensemble methods
   10 October: C16, C17

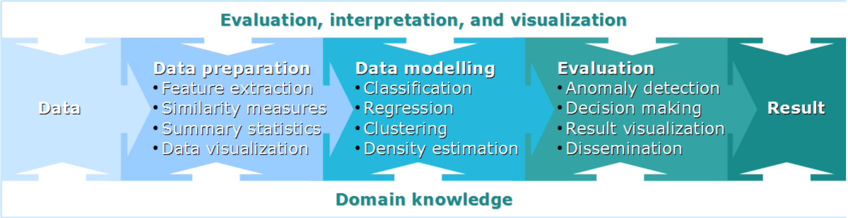
Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    10 October: C18
11. Mixture models and density estimation
    11 October: C19, C20
12. Association mining
    11 October: C21

Recap

13. Recap
    11 October: C1-C21

Piazza online help: `https://piazza.com/dtu.dk/fall2019/october2019`




## Learning Objectives

- Statistically evaluate cross-validation results
- Account for the assumptions made in Naïve Bayes
- Apply Bayes Theorem to obtain the class posterior likelihood

## Why test?

Statistical evaluation can mean a number of things:

- A social media company wish to know if introducing a new ad-placement method increases the click-through rate over another
- How many customers are likely click adds next month?
- How well can a neural network model learn to distinguish between diseased/non-diseased X-rays?
- Should I recommend that people use my neural network model over a competing method?

Tests can provide two things:

- An objective way to choose between methods
- A way to quantify model performance which takes uncertainty into account

# Outline: why not just one test?

- What is our overall **objective**? What conclusions do we want?
- What is our fundamental **evaluation criteria**?
- What specific test should I use? (classification, regression, etc.)

# The objective and evaluation criteria

- We compare models based on how well they **generalize to future data**
- Suppose we have data $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$ and two models $\mathcal{M}_A$, $\mathcal{M}_B$
- Training on $\mathcal{D}$, we obtain prediction rules

$$f_{\mathcal{D},A}, \quad \text{and} \quad f_{\mathcal{D},B}.$$

- Compared by the **difference in generalization error**:

$$z_{\mathcal{D}} = E^{\text{gen}}_{\mathcal{D},A} - E^{\text{gen}}_{\mathcal{D},B}$$

$$E^{\text{gen}}_{\mathcal{D},A} = \int p(\boldsymbol{x}, y) L(f_{\mathcal{D},A}(\boldsymbol{x}), y) d\boldsymbol{x} dy, \quad E^{\text{gen}}_{\mathcal{D},B} = \int p(\boldsymbol{x}, y) L(f_{\mathcal{D},B}(\boldsymbol{x}), y) d\boldsymbol{x} dy.$$

- If $z_{\mathcal{D}} < 0$, it means **that $\mathcal{M}_A$ is better than $\mathcal{M}_B$... ...when trained on $\mathcal{D}$**
- This is **one possible objective**:

Setup I Statistical tests of performance considering the **specific** training set $\mathcal{D}$?

## A more general **objective**

- Compared by the difference in generalization error:

$$z_{\mathcal{D}} = E^{\text{gen}}_{\mathcal{D},A} - E^{\text{gen}}_{\mathcal{D},B}$$

- Therefore, if you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for $\mathcal{D}'$ (from same distribution as $\mathcal{D}$)
- Therefore, our experiment is not independently reproducible
- To overcome this, test if $\mathcal{M}_A$ is better than $\mathcal{M}_B$ when averaging over dataset

$$z = \mathbb{E}_{\mathcal{D}}[z_{\mathcal{D}}] < 0$$

$$E^{\text{gen}} = \int \left[ \int L(f_{\mathcal{D}}(\boldsymbol{x}), y) p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} dy \right] p(\mathcal{D}) d\mathcal{D}$$

- If $z < 0$, it means **$\mathcal{M}_A$ is better than $\mathcal{M}_B$ ... on a typical training set**

Setup II *Statistical tests of performance considering a dataset of size* $N$

# Choices, choices

Setup I Statistical tests of performance considering the **specific** training set $\mathcal{D}$

Setup II *Statistical tests of performance considering* ***a dataset*** *of size* $N$

We cannot tell you what to do as it fundamentally depends on your situation and what you want to conclude. But write the conclusion correctly!

- Setup II is a more general (impressive) conclusion
- Setup II is probably what we want in science
- Setup II requires (a lot of) cross-validation
- If you have a single train/test split, use setup I

We will consider **setup I** here

## Statistical goals

Hypothesis testing Determine whether there is an effect, i.e. choose between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

Estimation Determine (likely) value $z \approx \hat{z}$ and an interval $[z_L, z_U]$ that likely contains $z$

- Focus should be on estimation: No two models are equal and a difference of $1\%$ is often of little interest
- Use hypothesis testing as a decision rule or to color entries in a table

## Connecting objective to numbers

- We want to draw conclusions about the difference in performance:

$$z_{\mathcal{D}} = E^{\text{gen}}_{\mathcal{D},A} - E^{\text{gen}}_{\mathcal{D},B}$$

$$E^{\text{gen}}_{\mathcal{D},A} = \int p(\boldsymbol{x}, y) L(f_{\mathcal{D},A}(\boldsymbol{x}), y) d\boldsymbol{x} dy, \quad E^{\text{gen}}_{\mathcal{D},B} = \int p(\boldsymbol{x}, y) L(f_{\mathcal{D},B}(\boldsymbol{x}), y) d\boldsymbol{x} dy.$$

- This can be estimated as

$$\begin{aligned} \hat{z}_{\mathcal{D}} &= \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} \left[ L(f_{\mathcal{D},A}(\boldsymbol{x}_i), y_i) - L(f_{\mathcal{D},B}(\boldsymbol{x}_i), y_i) \right] \\ &= \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} z_i, \quad \text{where:} \quad z_i = L(f_{\mathcal{D},A}(\boldsymbol{x}_i), y_i) - L(f_{\mathcal{D},B}(\boldsymbol{x}_i), y_i). \end{aligned}$$

## Abstracting to a statistical question

Consider data as the $n$ numbers

$$D = (z_1, \ldots, z_n). \tag{1}$$

General form of the problem: Draw conclusions about

$$\theta = E^{\text{gen}}_{A,\mathcal{D}} - E^{\text{gen}}_{B,\mathcal{D}}$$

Based on $D$ and the estimate:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} z_i. \tag{2}$$

# Statistical tools: Parameter

- We assume $z_i$ is a realization of a random variable $Z_i$
- It has density

$$p(Z_i = z_i|\theta) = p_\theta(z_i)$$

- Density of all dataset

$$p_\theta(D) = \prod_{i=1}^{n} p_\theta(z_i). \tag{3}$$

- Returning to our goals:
    - **estimating plausible ranges of $\theta$**
    - **hypothesis testing such as whether $\theta$ takes a particular value**
- Let's look at the statistical tools to accomplish this

# Statistical tools: Statistic and estimator

**Statistic** A statistic is a function of the data $D$ and will be denoted $t$. For instance, the mean and variance are both statistics:

$$t_0(D) = \frac{1}{n}\sum_{i=1}^{n} Z_i, \text{ or } t_1(D) = \frac{1}{n}\sum_{i=1}^{n}(Z_i - t_0(D))^2.$$

**Estimator** An estimator is a statistic $t$ of $D$ such that $t(D)$ is close to $\theta$. In the examples we will consider the mean

$$t_0(D) = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

## Statistical tools: Confidence interval

- A **confidence interval** (CI) is an interval $[\theta_L, \theta_U]$ which likely contains $\theta$
- The CI is a function of the data $D$. $\theta_L$ and $\theta_U$ are two statistics and for a concrete dataset the interval is computed to be

$$[\theta_L(D), \theta_U(D)]. \tag{4}$$

- With probability $1 - \alpha$, the true value $\theta$ should fall within the confidence interval $[\theta_L(D), \theta_U(D)]$ as we randomize over different datasets

$$P_\theta(\theta \in [\theta_L, \theta_U]) = 1 - \alpha. \tag{5}$$

## Statistical tools: Null hypothesis testing and $p$-value

- Determining whether a hypothesis $H_0$ about the parameters (the **null hypothesis**) is true or false

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0$$

- Intuitively, if $H_0$ is true, the data should behave in a certain way. **We test if the data is implausible assuming $H_0$**
- Specifically, let $t$ be a statistic, for our purpose

$$t(D) = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

  On our dataset it has a particular value $t_0 = \frac{1}{n} \sum_{i=1}^{n} z_i$
- We can compute the density $t(D)$ takes a particular value given $H_0$ is true:

$$p(t(D) = t | H_0) = p_{\theta=\theta_0}(t(D) = t)$$

- $p$-value is the chance $t(D)$ is at least as extreme as what we actually observed:

$$\textit{p-value}: \quad p = P\left(t(D) > |t_0| \mid H_0\right) = P_{\theta=\theta_0}(t(D) \geq |t_0|). \qquad (6)$$

## Setup I: Fixed training set

Suppose we carry out cross-validation to obtain:

$$(\mathcal{D}_1^{\text{train}}, \mathcal{D}_1^{\text{test}}), \ldots, (\mathcal{D}_K^{\text{train}}, \mathcal{D}_K^{\text{test}}). \tag{7}$$

We collect these into (paired) vectors of predictions and true values:

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{\boldsymbol{y}}_1 \\ \hat{\boldsymbol{y}}_2 \\ \vdots \\ \hat{\boldsymbol{y}}_K \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1^{\text{train}} \\ \boldsymbol{y}_2^{\text{train}} \\ \vdots \\ \boldsymbol{y}_K^{\text{train}} \end{bmatrix}. \tag{8}$$

## Evaluation of a single classifier

- Define:

$$c_i = \begin{cases} 1 & \text{if } \hat{y}_i = y_i \\ 0 & \text{if otherwise.} \end{cases}$$

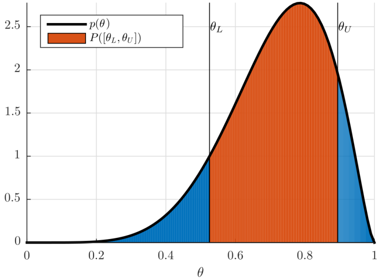
- Number of accurate guesses:

$$m = \sum_{i=1}^{n} c_i.$$

- Let the chance the classifier is correct be $\theta$. Then, from **Lecture 4**, we know

$$p(\theta|m,n) = \mathrm{Beta}(\theta|a,b), \quad a = m + \frac{1}{2}, \text{ and } b = n - m + \frac{1}{2}. \tag{9}$$

# Intermezzo: cummulative densities




- Consider a general probability density $p(\theta)$ of a parameter $\theta$
- Recall that by the definition of $p$, then

$$p(\theta \text{ in the interval } [\theta_L, \theta_U]) = p([\theta_L, \theta_U]) = \int_{\theta_L}^{\theta_U} p(\theta) d\theta$$

- Suppose $p([\theta_L, \theta_U]) = 0.95$.
  The interpretation is **we are nearly certain that $\theta$ is in $[\theta_L, \theta_U]$**.
- We can use this to define intervals that likely contain the true parameter

# Credibility interval

- We define the cummulative density function $\mathrm{cdf}$ as

$$\mathrm{cdf}(\theta) = P([-\infty, \theta]) = \int_{-\infty}^{\theta} p(\theta') d\theta'$$

- The blue area is therefore $P(A) = \mathrm{cdf}(\theta)$
- We can define the inverse of the cdf

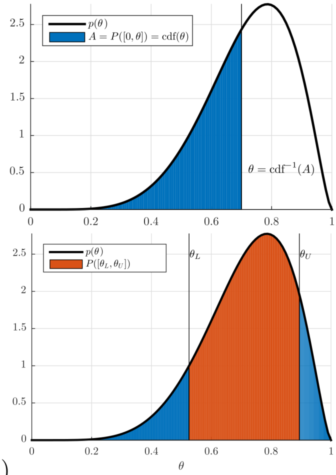
$$\theta = \mathrm{cdf}^{-1}(x), \quad x = p([-\infty, \theta])$$

- Therefore, the $1 - \alpha$ candidate confidence interval

$$\theta_L = \mathrm{cdf}^{-1}\left(\frac{\alpha}{2}\right), \ \mathrm{cdf}^{-1}\left(1 - \frac{\alpha}{2}\right)$$

In which case

$$\begin{aligned} P([\theta_L, \theta_U]) &= P([-\infty, \theta_U]) - P([-\infty, \theta_L]) \\ &= \mathrm{cdf}(\theta_U) - \mathrm{cdf}(\theta_L) \\ &= 1 - \alpha \end{aligned}$$



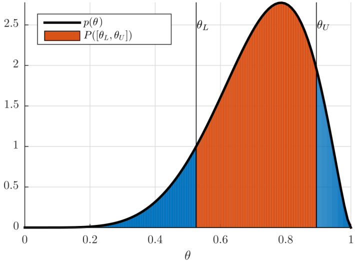
# Evaluating a single classifier

- If $m$ is the number of accurate guesses, then

$$p(\theta|m,n) = \mathrm{Beta}(\theta|a,b), \quad a = m + \frac{1}{2}, \text{ and } b = n - m + \frac{1}{2}.$$

- The $1-\alpha$ confidence interval is given as $[\theta_L, \theta_U]$:

$$\theta_L = \mathrm{cdf}_B^{-1}\left(\frac{\alpha}{2}|a,b\right) \text{ if } m > 0 \text{ otherwise } \theta_L = 0$$
$$\theta_U = \mathrm{cdf}_B^{-1}\left(1-\frac{\alpha}{2}|a,b\right) \text{ if } m < n \text{ otherwise } \theta_U = 1$$
$$\hat{\theta} = \mathbb{E}[\theta] = \frac{a}{a+b}$$



## Comparing two classifiers

- Assume we have predictions from both classifiers:

$$\hat{\boldsymbol{y}}^A = \hat{y}_1^A, \dots, \hat{y}_n^A, \quad \hat{\boldsymbol{y}}^B = \hat{y}_1^B, \dots, \hat{y}_n^B.$$

- As before, we want to know if the classifiers are correct or not:

$$c_i^A = \begin{cases} 1 & \text{if } \hat{y}_i^A = y_i \\ 0 & \text{if otherwise.} \end{cases} \quad \text{and} \quad c_i^B = \begin{cases} 1 & \text{if } \hat{y}_i^B = y_i \\ 0 & \text{if otherwise.} \end{cases}$$

- The relevant information is the contingency table:

$$n_{11} = \sum_{i=1}^{n} c_i^A c_i^B = \{\text{Both classifiers are correct}\}$$

$$n_{12} = \sum_{k=1}^{n} c_i^A (1 - c_i^B) = \{A \text{ is correct, } B \text{ is wrong}\}$$

$$n_{21} = \sum_{k=1}^{n} (1 - c_i^A) c_i^B = \{A \text{ is wrong, } B \text{ is correct}\}$$

$$n_{22} = \sum_{k=1}^{n} (1 - c_i^A)(1 - c_i^B) = \{\text{Both classifiers are wrong}\}$$

## Comparing two classifiers: McNemars test

- We want to compare the accuracy difference:

$$\theta = \theta_A - \theta_B$$

- It is possible to show (approximately)

$$\begin{aligned} p(\theta|\boldsymbol{n}) &= \frac{1}{2}\mathrm{Beta}\left(\frac{\theta+1}{2} \,\middle|\, \alpha = p, \beta = q\right) \\ \theta_L &= 2\mathrm{cdf}_B^{-1}\left(\frac{\alpha}{2} \,\middle|\, \alpha = p, \beta = q\right) - 1 \\ \theta_U &= 2\mathrm{cdf}_B^{-1}\left(1 - \frac{\alpha}{2} \,\middle|\, \alpha = p, \beta = q\right) - 1 \end{aligned}$$

$$\begin{aligned} p &= \frac{E_\theta + 1}{2}(Q-1) \\ q &= \frac{1 - E_\theta}{2}(Q-1) \\ E_\theta &= \frac{n_{12} - n_{21}}{n}, \quad Q = \frac{n^2(n+1)(E_\theta+1)(1-E_\theta)}{n(n_{12}+n_{21}) - (n_{12}-n_{21})^2} \end{aligned}$$

- For a $p$-value, note that $A$ is better than $B$ if $n_{12} > n_{21}$
- Chance of a particular value $n_{12}$ given $H_0$ is $p_{\text{binom}}(n_{12}|\theta = \frac{1}{2}, N = n_{12} + n_{21})$
- The probability of obtaining as extreme value as the one observed is:

$$\begin{aligned} p &= P(N_{12} \leq m|H_0) + P(N_{21} \leq m|H_0) \\ &= 2\mathrm{cdf}_{\text{binom}}\left(m = \min\{n_{12}, n_{21}\} \,\middle|\, \theta = \frac{1}{2}, N = n_{12} + n_{21}\right) \end{aligned}$$

# Confidence interval for a regression model

- Use cross-validation to obtain predictions $\hat{y}_i$ and true values $y_i$. Select loss

$$z_i = |\hat{y}_i - y_i| \quad \text{or} \quad z_i = (\hat{y}_i - y_i)^2 \tag{10}$$

- Estimated error is: $\hat{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$.
- Assume each error is normally distributed (**warning!**)

$$p(D|u,\sigma^2) = \prod_{i=1}^{n} \mathcal{N}(z_i|u,\sigma^2)$$

- It is possible to show $u$ follows a generalized Student's $t$-distribution:

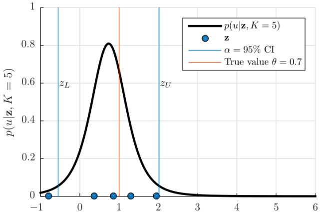
$$p(u|D) = p_{\mathcal{T}}(u|\nu = n-1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

with parameters $\hat{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ and $\tilde{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{(z_i - \hat{z})^2}{n(n-1)}}$.

- The Student's $t$-distribution has density

$$\textit{Student } t\textit{-distribution} \quad p_{\mathcal{T}}(x|\nu,\mu,\sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{1}{\nu}\left[\frac{x-\mu}{\sigma}\right]^2\right)^{-\frac{\nu+1}{2}}.$$

# Confidence interval for a regression model




- Step back: Assuming $z_i = L(y_i, \hat{y}_i)$ and

$$z_i \sim \mathcal{N}(z_i|\mu = u, \sigma^2)$$

- In this case $u$ is the average error. Since we have shown:

$$p(u|D) = p_{\mathcal{T}}(u|\nu = n-1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

- An approximate $1-\alpha$ confidence interval is:

$$z_L = \mathrm{cdf}_{\mathcal{T}}^{-1}\left(\frac{\alpha}{2} \,\middle|\, \nu, \hat{z}, \tilde{\sigma}\right), \; z_U = \mathrm{cdf}_{\mathcal{T}}^{-1}\left(1 - \frac{\alpha}{2} \,\middle|\, \nu, \hat{z}, \tilde{\sigma}\right). \tag{11}$$

## Comparing two regression models

- Use cross-validation to obtain (paired) predictions along with true values $y_i$

$$\hat{y}_1^A, \dots, \hat{y}_n^A, \quad \text{and} \quad \hat{y}_1^B, \dots, \hat{y}_n^B. \tag{12}$$

- Select a loss-function to compute the per-observation losses as in

$$z_1^A, \dots, z_n^A, \quad \text{and} \quad z_1^B, \dots, z_n^B.$$

- Note that

$$\begin{aligned} z = E^{\text{gen}}_{A,\mathcal{D}} - E^{\text{gen}}_{B,\mathcal{D}} \approx \hat{z} &= \left( \frac{1}{n} \sum_{i=1}^{n} z_i^A \right) - \left( \frac{1}{n} \sum_{i=1}^{n} z_i^B \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} z_i, \quad \text{where } z_i = z_i^A - z_i^B \end{aligned}$$

Compute a $1 - \alpha$ CI using methods on previous slide

## Comparing two regression models: $p$-values

- Still using

$$z = E_A^{\text{gen}} - E_B^{\text{gen}} \approx \hat{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, \qquad \text{where } z_i = z_i^A - z_i^B$$

- Assuming

$$z_i \sim \mathcal{N}(z_i|\mu = u, \sigma^2)$$

- where $u$ is the true difference in error function we have shown:

$$p(u|D) = p_{\mathcal{T}}(u|\nu = n-1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

- Therefore, we can test the hypothesis

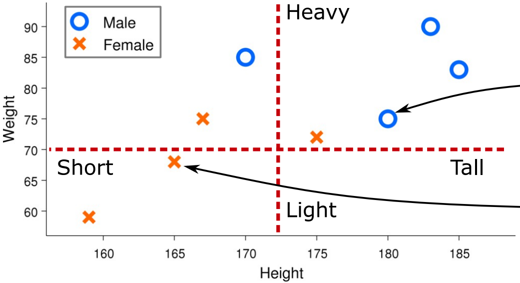
$$H_0 : \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have the same performance, } u = 0 \tag{13}$$

$$H_1 : \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have different performance, } u \neq 0. \tag{14}$$

- A $p$-value can be computed as

$$\begin{aligned} p &= P\left(Z \geq |\hat{z}| \mid H_0\right) = 2\int_{-\infty}^{-|\hat{z}|} p_{\mathcal{T}}\left(z \mid \nu = n-1, \mu = 0, \sigma = \tilde{\sigma}\right) dz \\ &= 2\mathrm{cdf}_{\mathcal{T}}\left(-|\hat{z}| \mid \nu = n-1, \mu = 0, \sigma = \tilde{\sigma}\right). \end{aligned} \tag{15}$$

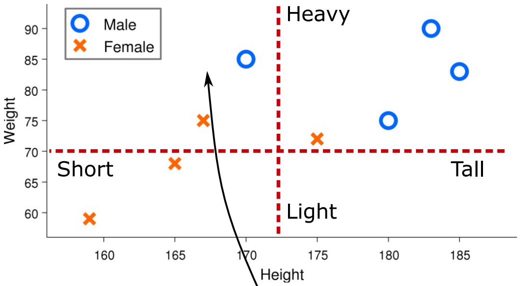
# Bayes and Naive-Bayes




| Gender $y$ | Tall $x_1$ | Heavy $x_2$ |
|---|---|---|
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |

$$p(y|x_1, x_2) = \frac{p(x_1, x_2|y)p(y)}{\sum_{k=0}^{1} p(x_1, x_2|y = k)p(y = k)}$$

# Example 1: Normal Bayes




| Gender $y$ | Tall $x_1$ | Heavy $x_2$ |
|---|---|---|
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |

Probability a short, heavy person is male:

$$P(y=0|x_1=0,x_2=1) = \frac{p(x_1=0,x_2=1|y=0)p(y=0)}{\sum_{k=0}^{1} p(x_1=0,x_2=1|y=k)p(y=k)}$$

## Example 1: Solution

Probability a short, heavy person is male:

$$P(y=0|x_1=0,x_2=1) = \frac{p(x_1=0,x_2=1|y=0)p(y=0)}{\sum_{k=0}^{1} p(x_1=0,x_2=1|y=k)p(y=k)}$$
$$= \frac{\frac{1}{4}\frac{4}{8}}{\frac{1}{4}\frac{4}{8}+\frac{1}{4}\frac{4}{8}} = \frac{1}{2}$$

## A practical problem with Bayesian classifier

- In general:

$$p(y|x_1, x_2, \ldots, x_M) = \frac{p(x_1, x_2, \ldots, x_M|y)p(y)}{\sum_{k=0}^{K-1} p(x_1, x_2, \ldots, x_M|y=k)p(y=k)}$$

$$p(x_1, \ldots, x_M|y=k) = \frac{\text{Nr. obs where } y=k \text{ and we measure } x_1, \ldots, x_M}{\text{Observations where } y=k}$$

## A practical problem with Bayesian classifier

- In general:

$$p(y|x_1, x_2, \ldots, x_M) = \frac{p(x_1, x_2, \ldots, x_M|y)p(y)}{\sum_{k=0}^{K-1} p(x_1, x_2, \ldots, x_M|y=k)p(y=k)}$$

$$p(x_1, \ldots, x_M|y=k) = \frac{\text{Nr. obs where } y=k \text{ and we measure } x_1, \ldots, x_M}{\text{Observations where } y=k}$$

- Naive Bayes assumption

$$p(x_1, x_2, \ldots, x_M|y) = p(x_1|y)p(x_2|y) \times \cdots \times p(x_M|y)$$

## A practical problem with Bayesian classifier

- In general:

$$p(y|x_1, x_2, \ldots, x_M) = \frac{p(x_1, x_2, \ldots, x_M|y)p(y)}{\sum_{k=0}^{K-1} p(x_1, x_2, \ldots, x_M|y=k)p(y=k)}$$

$$p(x_1, \ldots, x_M|y=k) = \frac{\text{Nr. obs where } y=k \text{ and we measure } x_1, \ldots, x_M}{\text{Observations where } y=k}$$

- Naive Bayes assumption

$$p(x_1, x_2, \ldots, x_M|y) = p(x_1|y)p(x_2|y) \times \cdots \times p(x_M|y)$$

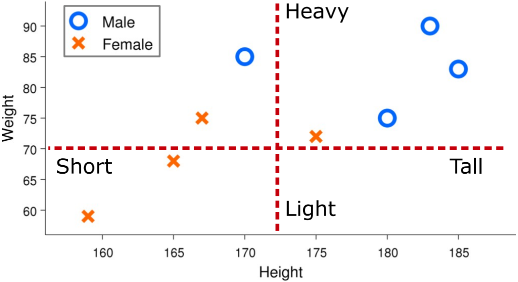
- Naive Bayes classifier

$$\begin{aligned} p(y|x_1, x_2, \ldots, x_M) &= \frac{p(x_1, x_2, \ldots, x_M|y)p(y)}{\sum_{k=0}^{1} p(x_1, x_2, \ldots, x_M|y=k)p(y=k)} \\ &= \frac{p(x_1|y)p(x_2|y) \times \cdots \times p(x_M|y)p(y)}{\sum_{k=0}^{1} p(x_1|y=k)p(x_2|y=k) \times \cdots \times p(x_M|y=k)p(y=k)} \end{aligned}$$

## Example 2:

- Naive Bayes classifier (Probability someone is a female given they are heavy and tall)

$$p(y=1|x_1=1,x_2=1)=\frac{p(x_1|y)p(x_2|y)p(y)}{\sum_{k=0}^{1}p(x_1|y=k)p(x_2|y=k)p(y=k)}$$




| Gender $y$ | Tall $x_1$ | Heavy $x_2$ |
|---|---|---|
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |

## Example 2: Solution

- Naive Bayes classifier (Probability someone is a female given they are heavy and tall)

$$
\begin{aligned}
p(y=1|x_1=1,x_2=1) &= \frac{p(x_1|y)p(x_2|y)p(y)}{\sum_{k=0}^{1} p(x_1|y=k)p(x_2|y=k)p(y=k)} \\
&= \frac{\frac{1}{4}\frac{2}{4}\frac{1}{2}}{\frac{1}{4}\frac{2}{4}\frac{1}{2}+\frac{3}{4}\frac{4}{4}\frac{1}{2}} = \frac{2}{2+12} = \frac{1}{7}
\end{aligned}
$$

## Quiz 1, Naive-Bayes (Spring 2012)

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| P2 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| P3 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| P4 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| P5 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| P6 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |

Table 1: Table indicating whether 10 songs denoted S1–S10 are downloaded to 6 different phones denoted P1–P6. P1 and P2 given in red are phones that belong to females whereas P3, P4, P5, and P6 given in blue belong to males.

The phones P1 and P2 are owned by females whereas P3, P4, P5 and P6 are owned by males (this is indicated in red and blue respectively in Table 1). We would like to predict whether a phone is owned by a male based on whether or not the songs S1, S2 and S3 have been downloaded. We will therefore classify whether the phone belongs to a male or female considering only the attributes S1, S2 and S3 and the data in Table 1. We will apply a Naïve Bayes classifier that assumes independence between these attributes. Given that a phone has installed songs 1, 2 and 3 (i.e., S1=1, S2=1 and S3=1) What is the probability that the phone is owned by a male according to the Naïve Bayes classifier?

A. 1/12

B. 1/6

C. 2/3

D. 1

E. Don't know.

$$p(y|x_1, x_2, \ldots, x_M) = \frac{p(x_1|y) \times \cdots \times p(x_M|y)p(y)}{\sum_{k=0}^{1} p(x_1|y=k) \times \cdots \times p(x_M|y=k)p(y=k)}$$

According to the Naïve Bayes classifier we have

$$P(Male|S1=1,S2=1,S3=1) =$$

$$\frac{\begin{pmatrix} P(S1=1|Male)\times \\ P(S2=1|Male)\times \\ P(S3=1|Male)\times \\ P(Male) \end{pmatrix}}{\begin{pmatrix} P(S1=1|Female)\times \\ P(S2=1|Female)\times \\ P(S3=1|Female)\times \\ P(Female) \end{pmatrix} + \begin{pmatrix} P(S1=1|Male)\times \\ P(S2=1|Male)\times \\ P(S3=1|Male)\times \\ P(Male) \end{pmatrix}}$$

$$= \frac{2/4 \cdot 2/4 \cdot 2/4 \cdot 4/6}{2/2 \cdot 0/2 \cdot 1/2 \cdot 2/6 + 2/4 \cdot 2/4 \cdot 2/4 \cdot 4/6} = 1.$$

# Robust estimation

- Probability of y given x for discrete variables

$$p(y|x) = \frac{n_c}{n}$$

Number of objects having value *y and x*

Total number of objects that have value x

  - Not defined when *n*=0

- Robust estimation

$$p(y|x) = \frac{n_c + m_c}{n + m}$$

Pseudo observations of objects having value y and x

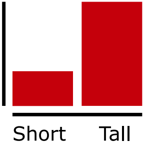
Equivalent pseudo-sample size of objects having value x

  - If no objects take value *x* the probability will be $\frac{m_c}{m}$
  - Corresponds to putting *m* extra objects into the data set

## Bayesian classifiers

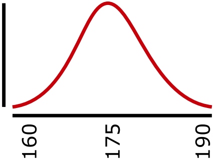
$p(\mathrm{Height}|\mathrm{Gender} = \mathrm{Male})$

- Handling continuous attributes
  - Two way split (x<a)
    - Converts into binary attribute
      (We have used this in the previous example)




  - Discretize into a number of bins
    - Converts into discrete ordinal attribute


  - Probability density estimation
    - Assume attribute follows a Normal distribution
    - Use data to compute parameters
      (mean and variance)



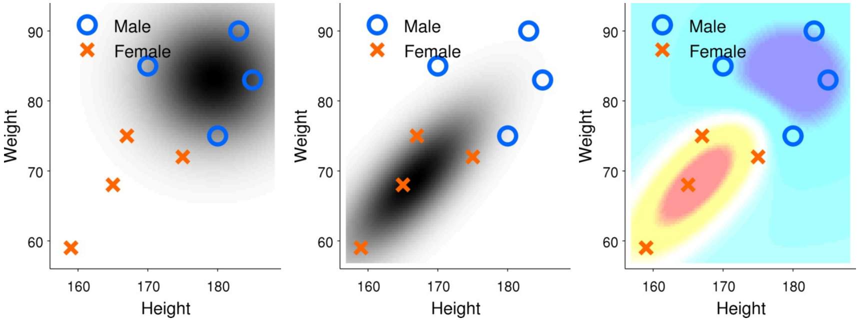
# Bayesian classification by the multivariate normal distribution

Continuous density estimation

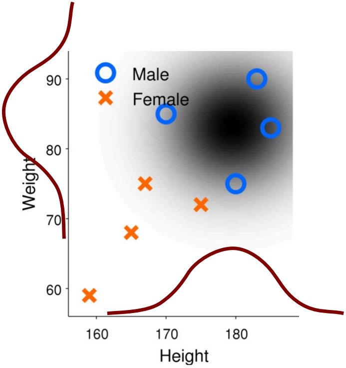
- Fit a Normal distribution to each class
  - Compute class mean and covariance
- Classify using Bayes rule as before

$$P(\boldsymbol{x}|y=c) = \frac{1}{(2\pi)^{M/2} det(\boldsymbol{\Sigma}_c)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)\right)$$

$$P(y=c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y=c)P(y=c)}{\sum_{c'} P(\boldsymbol{x}|y=c')P(y=c')}$$



- What does the Naive Bayes assumption of independence of the attributes correspond to in terms of the parameters of the multivariate normal distribution?



## Resources

https://www.youtube.com Video explaining Naive Bayes

(https://www.youtube.com/watch?v=8yvBqhm92xA)

https://machinelearningmastery.com Statistical comparison of the cross-validation estimate of the generalization error is not a solved problem. This reference provides an overview of various issues and proposed solutions. Note no simple solution exists.

(https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/)

https://link.springer.com An arguably better (but slightly more complicated) way to compare the generalization error estimated from cross-validation. Note the method can be seen as an extension to the credibility-interval method presented here (https://link.springer.com/article/10.1007/s10994-015-5486-z)