DTU

02450 Introduction to Machine Learning and Data Mining

Week 7: Performance evaluation, Bayes, and Naive Bayes

Georgios Arvanitidis

18 March 2025

DTU Compute, Technical University of Denmark

Today

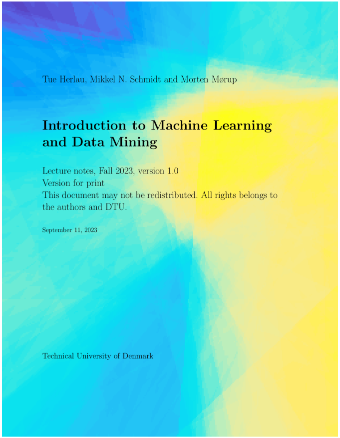
Feedback Groups of the day:

Zhentao Wei, Caroline Sofie Nortoft Dybdahl, Aatharva Ravindra Kawade, Amanda Elisabet Hietala, Frederik Lauritz Udby, Anna Victoria Vejlsby, Chris Lam Dao, Kasper Sebastian Kjølby Carlsen, Jonas Hvidtfeldt Møller, Eleni Karanikola, Nathalie Marie Hauch, Kasper Bo Christensen, Sofie Øksnes Bruaset, Ly Nguyen Hansen, Vinit Nilesh Vasa, Jan Steffen Heibel, Thomas Møller Pedersen, Ali Asadighafari, Connor Peter Rudd, Gustav Bellaiche, Weinan Yan, Jan-Dirk Randewijk, Ryan Mak, Aali Fida, Ralph Valentin, Bendik Barlinn Kjelstad, Casper Langthjem Bertelsen, Mads Peter Lysholm Jønsson, Johannes Nicolás Wildfeuer, Ro Nanak Prasad Lacoul, Niklas Julian Gerds, Anna Luna Svensson, Jannes Müller, Mayisha Maliha, Akhil Karnati, Trine Søgaard, Karítas Ósk Pálmadóttir, Jonathan Josua Pedersen Gammelgaard, Balint Norbert, Kristian Juul Rasmussen, Andrea De Pascale, Nina Schwalb

Reading/homework material:

Chapter 11, 13

P12.2, P13.1, P13.2



Lecture Schedule

1. Introduction
   4 February: C1,C2

Data: Feature extraction, and visualization

2. Summary statistics, similarity and visualization
   11 February: C4,C7
3. Computational linear algebra and PCA
   18 February: C3
4. Probability and probability densities
   25 February: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   4 March: C8, C9 **(Project 1 due 6 March at 17:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   11 March: C10, C12
7. **Performance evaluation, Bayes, and Naive Bayes**
   18 March: C11, C13
8. Artificial Neural Networks and Bias/Variance
   25 March: C14, C15
9. AUC and ensemble methods
   1 April: C16, C17

Unsupervised learning: Clustering and density estimation

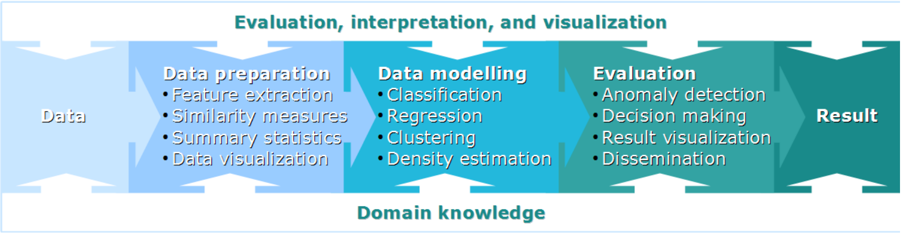
10. K-means and hierarchical clustering
    8 April: C18 **(Project 2 due 10 April at 17:00)**
11. Mixture models and density estimation
    22 April: C19, C20
12. Association mining
    29 April: C21

Recap

13. Recap and discussion of the exam
    6 May: C1-C21

Online help: Piazza
Videos of lectures: `https://panopto.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)

Learning Objectives



Learning Objectives

- Understand the two different evaluation setups
- Apply appropriate statistical tests to evaluate and compare models
- Account for the assumptions made in Naïve Bayes
- Apply Bayes theorem to obtrain the class posterior likelihood

Statistical testing

- A social media company wish to know if a new ad-placement method increases the click-through rate
- 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 my neural network model over a competing method?

All involve induction beyond the dataset

Statistical testing

Tests **can** provide:

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

Statistical testing

Tests **can** provide:

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

Tests **do not** provide

- Certain conclusions (Model A is better than B)
- A black-box recipe

Statistical testing

Tests **can** provide:

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

Tests **do not** provide

- Certain conclusions (Model A is better than B)
- A black-box recipe

Use statistical tests to aid your interpretation of your results not as an argument in itself

Outline

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

The objective and evaluation criteria

- Models are compared 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

$$f_{\mathcal{D},A} : \boldsymbol{x} \to y, \quad \text{and} \quad f_{\mathcal{D},B} : \boldsymbol{x} \to y.$$

The objective and evaluation criteria

- Models are compared 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

$$f_{\mathcal{D},A} : \boldsymbol{x} \to y, \quad \text{and} \quad f_{\mathcal{D},B} : \boldsymbol{x} \to y.$$

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

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

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

The objective and evaluation criteria

- Models are compared 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

$$f_{\mathcal{D},A} : \boldsymbol{x} \to y, \quad \text{and} \quad f_{\mathcal{D},B} : \boldsymbol{x} \to y.$$

- 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 $\mathcal{M}_B$...**

The objective and evaluation criteria

- Models are compared 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

$$f_{\mathcal{D},A} : \boldsymbol{x} \to y, \quad \text{and} \quad f_{\mathcal{D},B} : \boldsymbol{x} \to y.$$

- 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 $\mathcal{D}$**

The objective and evaluation criteria

- Models are compared 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

$$f_{\mathcal{D},A} : \boldsymbol{x} \to y, \quad \text{and} \quad f_{\mathcal{D},B} : \boldsymbol{x} \to y.$$

- 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 $\mathcal{D}$**

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

An alternative objective

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

- If you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for an independent dataset $\mathcal{D}'$

An alternative objective

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

- If you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for an independent dataset $\mathcal{D}'$
- Therefore, the conclusion is not independently reproducible

An alternative objective

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

- If you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for an independent dataset $\mathcal{D}'$
- Therefore, the conclusion is not independently reproducible
- To overcome this, test if $\mathcal{M}_A$ is better than $\mathcal{M}_B$ when averaging over dataset

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

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

An alternative objective

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

- If you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for an independent dataset $\mathcal{D}'$
- Therefore, the conclusion is not independently reproducible
- To overcome this, test if $\mathcal{M}_A$ is better than $\mathcal{M}_B$ when averaging over dataset

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

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

- If $z < 0$, it means $\mathcal{M}_A$ **is better than** $\mathcal{M}_B$ **... using 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

Which to choose fundamentally depends on what you want to conclude

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

Which to choose fundamentally depends on what you want to conclude

- 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 tasks and tools

Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

Statistical tasks and tools

Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

Hypothesis testing Determine **whether there is an effect** by choosing between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

Statistical tasks and tools

Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

Hypothesis testing: Determine **whether there is an effect** by choosing between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

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

Statistical tasks and tools

Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

Hypothesis testing Determine **whether there is an effect** by choosing between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

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

- Evidence against H_0 is measured by a p**-value** (low p is evidence for an effect $z \neq 0$)

Statistical tasks and tools

Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

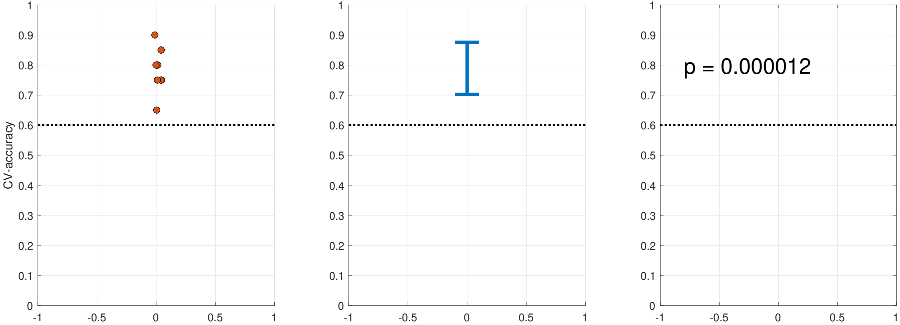
Hypothesis testing: Determine **whether there is an effect** by choosing between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

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

- Evidence against H_0 is measured by a **p-value** (low p is evidence for an effect $z \neq 0$)
- Estimation of $[z_L, z_U]$ done using an **α-confidence interval** (lower α means a more conservative, wider, interval)

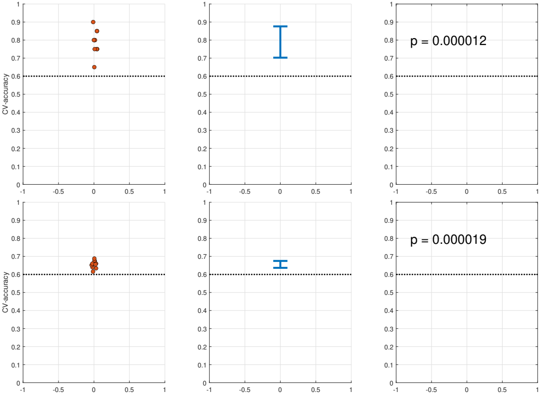
Choosing the right tool

- Consider binary classification using $N = 200$ samples
- We estimate test error using $K = 10$-fold CV (10 test-error estimates)
- Question: Is accuracy $E_A^{\text{gen}} \approx \frac{1}{K}\sum_{k=1}^{K} E_i^{\text{test}}$ greater than baseline θ_0?
- (Baseline classify everything as maximum class, accuracy 60%)



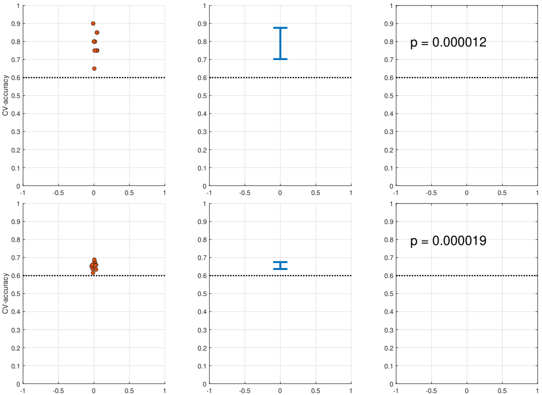
Which tool to use

- Top: $N = 200$ sample example
- Bottom: Harder problem using $N = 2000$ samples



Which tool to use

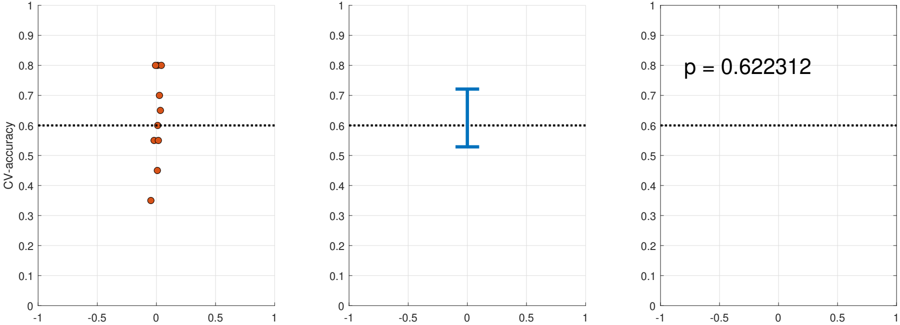
- Top: $N = 200$ sample example
- Bottom: Harder problem using $N = 2000$ samples



- p-value primarily measure of sample size (not **effect size!**)
- Which do **you** think are more informative?

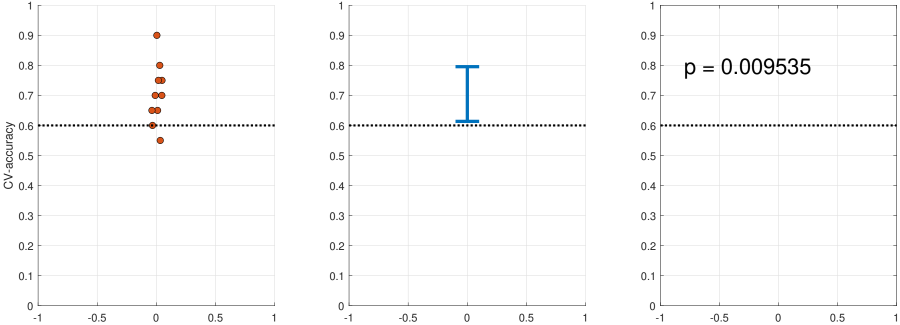
Variability

- New problem using $N = 200$ samples. Is there an effect?

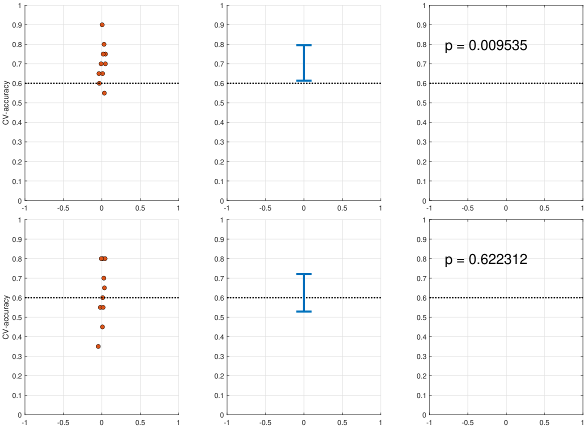


Variability

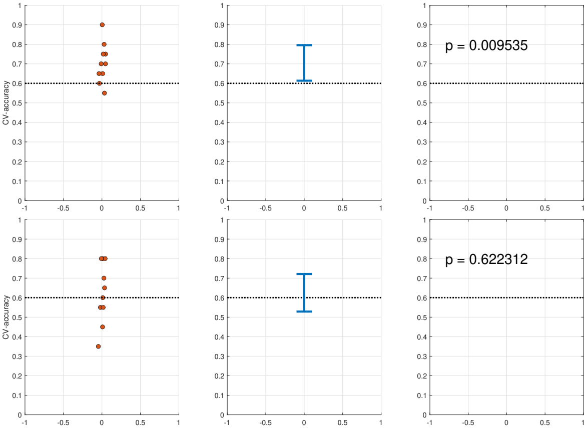
- Another problem using $N = 200$ samples. Is there an effect?



The nasty bit



The nasty bit



- Only difference is random variability in dataset
- Low p-value does **not necessarily** mean reproducible
  - Training many models will lead to false positives
  - Statistics **will not** fix unclear results; probably just lead to false positives

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

$$\hat{z}_{\mathcal{D}} = \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} \left[L(f_{\mathcal{D},A}(\boldsymbol{x}_i), y_i) - L(f_{\mathcal{D},B}(\boldsymbol{x}_i), y_i) \right]$$

$$= \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} z_i, \quad \text{where:} \quad z_i = L(f_{\mathcal{D},A}(\boldsymbol{x}_i), y_i) - L(f_{\mathcal{D},B}(\boldsymbol{x}_i), y_i).$$

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

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

Based on the estimate:

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

Statistical tools: Parameter

- 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

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

Statistical tools: Parameter

- 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 θ**
  - **hypothesis testing such as whether θ 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:

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

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 θ. 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 θ

Statistical tools: Confidence interval

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

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

Statistical tools: Confidence interval

- A **confidence interval** (CI) is an interval $[\theta_L, \theta_U]$ which likely contains θ
- The CI is a function of the data D. θ_L and θ_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 θ 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 **null hypothesis** H_0 about the parameters 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**

Statistical tools: Null hypothesis testing and p-value

- Determining whether a **null hypothesis** H_0 about the parameters 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:

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

Statistical tools: Null hypothesis testing and p-value

- Determining whether a **null hypothesis** H_0 about the parameters 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:

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

Setup I: Fixed training set

Suppose we carry out cross-validation to obtain:

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

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

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

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:

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

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 θ. Then, from **Lecture 4**, we know

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

Evaluating a single classifier (Jeffreys interval)

- If m is the number of accurate guesses, then

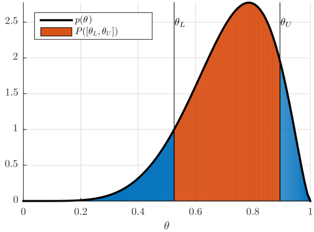
$$p(\theta|m,n) = \mathrm{Beta}(\theta|a,b), \quad a = m + \frac{1}{2}, \text{ and } b = n - m + \frac{1}{2}.$$

- 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:

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

Comparing two classifiers

- Assume we have predictions from both classifiers:

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

- 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:

$$\begin{aligned}
n_{11} &= \sum_{i=1}^{n} c_i^A c_i^B &&= \{\text{Both classifiers are correct}\} \\
n_{12} &= \sum_{k=1}^{n} c_i^A (1 - c_i^B) &&= \{A \text{ is correct, } B \text{ is wrong}\} \\
n_{21} &= \sum_{k=1}^{n} (1 - c_i^A) c_i^B &&= \{A \text{ is wrong, } B \text{ is correct}\} \\
n_{22} &= \sum_{k=1}^{n} (1 - c_i^A)(1 - c_i^B) &&= \{\text{Both classifiers are wrong}\}
\end{aligned}$$

Comparing two classifiers: McNemar's test

- We want to compare the accuracy difference: $\theta = \theta_A - \theta_B$
- It is possible to show (approximately)

$$
\begin{aligned}
p(\theta|\boldsymbol{n}) &= \frac{1}{2}\mathrm{Beta}\left(\frac{\theta+1}{2} \,\middle|\, a=f, b=g\right), \\
f &= \frac{E_\theta+1}{2}(Q-1) \quad g = \frac{1-E_\theta}{2}(Q-1) \\
E_\theta &= \frac{n_{12}-n_{21}}{n}, \quad Q = \frac{n^2(n+1)(E_\theta+1)(1-E_\theta)}{n(n_{12}+n_{21})-(n_{12}-n_{21})^2}. \\
\theta_L = 2\mathrm{cdf}_B^{-1}&\left(\frac{\alpha}{2} \,\middle|\, a=f, b=g\right)-1, \quad \theta_U = 2\mathrm{cdf}_B^{-1}\left(1-\frac{\alpha}{2} \,\middle|\, a=f, b=g\right)-1
\end{aligned}
\tag{10}
$$

- For a p-value, note that A is better than B if $n_{12} > n_{21}$
- A p-value can be obtained as:

$$
p = 2\mathrm{cdf}_{\mathrm{binom}}\left(m = \min\{n_{12}, n_{21}\} \,\middle|\, \theta = \frac{1}{2}, N = n_{12}+n_{21}\right)
$$

Confidence interval for a regression model

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

$$z_i = |\hat{y}_i - y_i| \quad \text{or} \quad z_i = (\hat{y}_i - y_i)^2 \tag{11}$$

- 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:

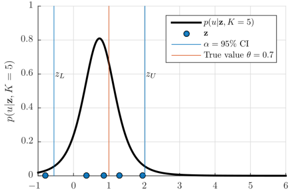
$$p(u|D) = p_{\mathcal{T}}(u|\nu = n-1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

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

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

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 rate. 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:

$$z_L = \mathrm{cdf}_{\mathcal{T}}^{-1}\left(\frac{\alpha}{2} \,\middle|\, \nu, \hat{z}, \tilde{\sigma}\right), \; z_U = \mathrm{cdf}_{\mathcal{T}}^{-1}\left(1 - \frac{\alpha}{2} \,\middle|\, \nu, \hat{z}, \tilde{\sigma}\right). \qquad (12)$$

Comparing two regression models

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

$$\hat{y}_1^A, \dots, \hat{y}_n^A, \quad \text{and} \quad \hat{y}_1^B, \dots, \hat{y}_n^B. \tag{13}$$

- 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

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

- Assume $z_i \sim \mathcal{N}(z_i|\mu = u, \sigma^2)$
- Compute a $1-\alpha$ CI using methods on previous slide

Comparing two regression models: p-values

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

- 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
$$H_0 : \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have the same performance, } u = 0 \tag{14}$$
$$H_1 : \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have different performance, } u \neq 0. \tag{15}$$
- A p-value can be computed as
$$p = 2\text{cdf}_{\mathcal{T}}\left(-|\hat{z}| \mid \nu = n-1, \mu = 0, \sigma = \tilde{\sigma}\right). \tag{16}$$

Which type of cross-validation?

- When using **setup I** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed

Which type of cross-validation?

- When using **setup I** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed
- Results will be significant with enough data

Which type of cross-validation?

- When using **setup I** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed
- Results will be significant with enough data
  - Focus on estimated effect size

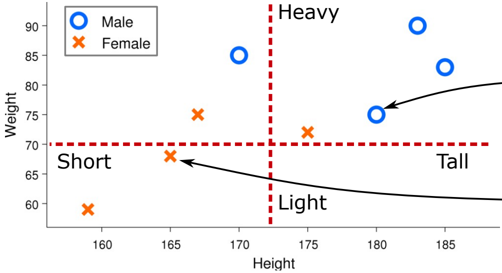
Which type of cross-validation?

- When using **setup I** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed
- Results will be significant with enough data
  - Focus on estimated effect size
  - Multiple-comparison problem

Which type of cross-validation?

- When using **setup I** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed
- Results will be significant with enough data
  - Focus on estimated effect size
  - Multiple-comparison problem
  - **Transparency, availability of datasets/code, breadth of testing, self-criticism** guarantees reprehensibility, not a sophisticated test
- In **setup II**, correlation of training data is taken into account and K-fold is optimal
  - Your **setup I** results do not generalize beyond your training data

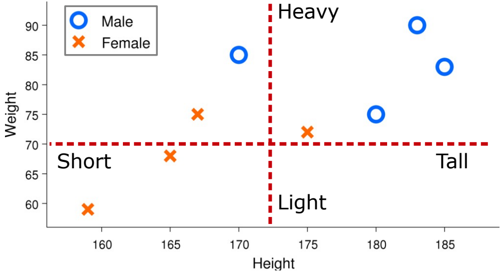
Bayes and Naive-Bayes



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:

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

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

- Naive Bayes assumption

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

- Naive Bayes classifier

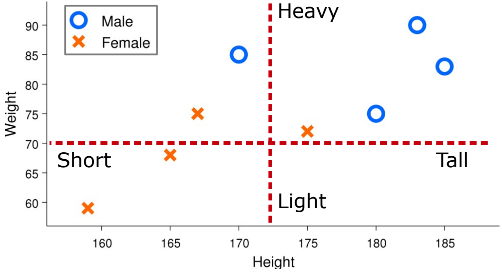
$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1, x_2, \dots, x_M|y)p(y)}{\sum_{k=0}^{1} p(x_1, x_2, \dots, x_M|y=k)p(y=k)}$$

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

Example 2: Solution

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

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



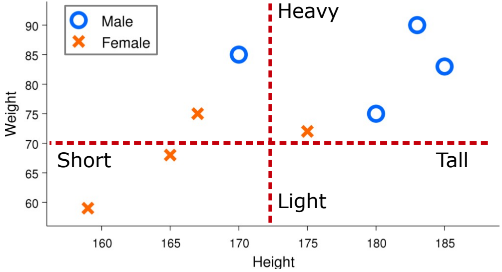
Gender y	Tall x_1	Heavy x_2
0	1	1
0	1	1
0	1	1
0	0	1
1	1	1
1	0	1
1	0	0
1	0	0

Example 2: Solution

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

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

$$= \frac{\frac{1}{4}\frac{2}{4}\frac{1}{2}}{\frac{1}{4}\frac{2}{4}\frac{1}{2} + \frac{3}{4}\frac{4}{4}\frac{1}{2}} = \frac{2}{2+12} = \frac{1}{7}$$



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

Quiz 1, Naive-Bayes (Spring 2012)

		Diabetic				Normal			
		$D1$	$D2$	$D3$	$D4$	$N1$	$N2$	$N3$	$N4$
Diabetic	$D1$	0	58.5	51.6	18.1	38.0	52.5	71.7	50.7
	$D2$	58.5	0	32.1	72.6	50.5	65.0	13.2	63.8
	$D3$	51.6	32.1	0	60.5	28.4	32.9	45.3	56.3
	$D4$	18.1	72.6	60.5	0	45.9	60.4	79.8	56.8
Normal	$N1$	38.0	50.5	28.4	45.9	0	17.5	63.7	50.7
	$N2$	52.5	65.0	32.9	60.4	17.5	0	78.2	57.2
	$N3$	71.7	13.2	45.3	79.8	63.7	78.2	0	71.0
	$N4$	50.7	63.8	56.3	56.8	50.7	57.2	71.0	0

The figure shows the distance between the first four diabetic (D1–D4) and normal (N1–N4) women. What are the number of misclassified observations for leave-one-out cross validation based on 3-nearest neighbor classification when only considering the 8 observations (i.e., D1–D4 and N1–N4) in the figure?

A. None of the observations will be misclassified.

B. 2 of the observations will be misclassified.

C. 6 of the observations will be misclassified.

D. All of the observations will be misclassified.

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

Robust estimation and non-binary data

Assume $p(x_1, \ldots, x_M|y) = \prod_{k=1}^{M} p(x_k|y)$

Defining $n_{x_j=k|y=c} = \sum_{i=1}^{N} \delta_{X_{ij},k}\delta_{y,c}$ we have more generally:

Binary case: $$p(x_j = 1|y = c) = \frac{n_{x_j=1|y=c} + \alpha}{N_c + 2\alpha}.$$

Categorical case: $$p(x_j = k|y = c) = \frac{n_{x_j=k|y=c} + \alpha}{N_c + K\alpha}.$$

Continious case: $$p(x_j = x|y = c) = \mathcal{N}(x|\mu = \mu_{j|c}, \sigma^2 = (\sigma_{j|c} + \alpha)^2)$$

$$\mu_{j|c} = \mathbb{E}_{y=c}[x_j] = \frac{1}{N_c}\sum_{i=1}^{N} \delta_{y_i,c} X_{ij},$$

$$\sigma_{j|c} = \hat{\mathrm{std}}_{y=c}[x_j] = \sqrt{\frac{1}{N_c - 1}\sum_{i=1}^{N} \delta_{y_i,c}(X_{ij} - \mu_c)^2}$$

Select these parameters using cross-validation.

Bayesian classification by the multivariate normal distribution

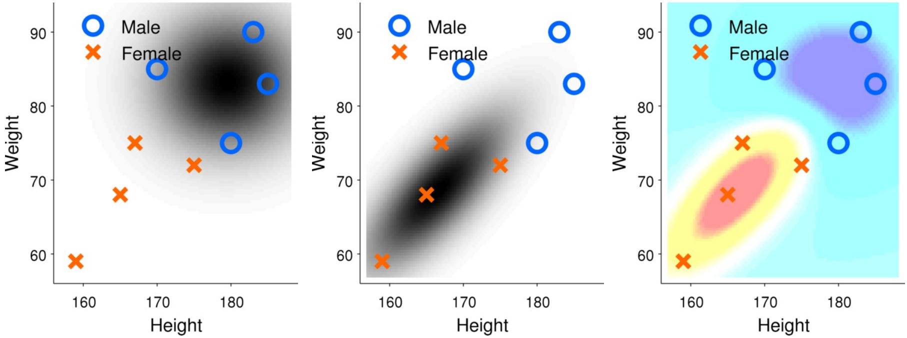
$$P(\boldsymbol{x}|y=c) = \frac{1}{(2\pi)^{\frac{M}{2}} det(\boldsymbol{\Sigma}_c)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)\right)$$

Continuous density estimation

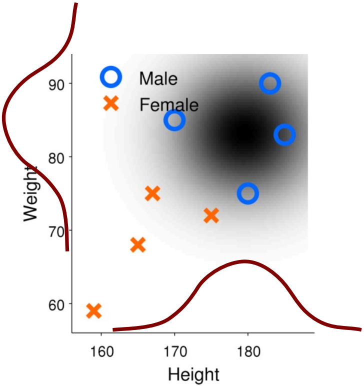
- Fit a Normal distribution to each class.
  - Compute class mean and covariance

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

- Classify using Bayes rules as before



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



Midterm practice test

Look at the test on DTU Learn. Note the test is not part of your evaluation.

Midterm question 1

In the analysis of house prices the following attributes were collected for a house: The year the house was built (denoted YEAR), the size of the house given in square meters (denoted SIZE) the county in which the house is located (denoted LOCATION). Which statement about the three attributes is correct?

A. YEAR is ratio, SIZE is interval and LOCATION is nominal

B. YEAR is interval, SIZE is ratio and LOCATION is nominal

C. YEAR is interval, SIZE is ratio and LOCATION is ordinal

D. YEAR is interval, SIZE is ratio and LOCATION is interval

E. Don't know.

Midterm question 2

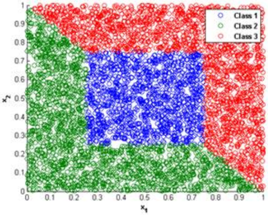
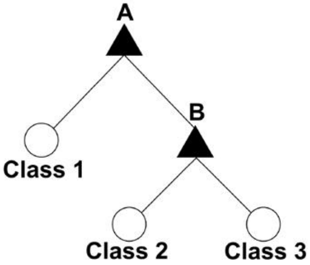


Figure 1



Consider the classification problem given in figure 1 and the Decision Tree shown below it with two decision nodes denoted A and B. We will let $\boldsymbol{x}_n = (x, y)$ denote a 2-dimensional observation such that $\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}$ denotes the subtraction of 0.5 from each of the two coordinates of $\boldsymbol{x}_n$. Which one of the following classification rules would lead to a correct classification of the data?

A. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B : $\|\boldsymbol{x}_n\|_\infty \leq 1$

B. A: $\|\boldsymbol{x}_n\|_1 \leq 1$, B: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$

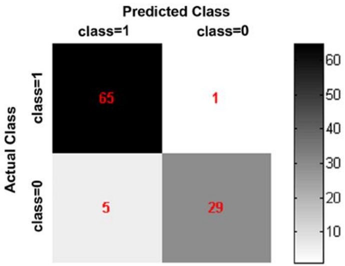
C. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B : $\|\boldsymbol{x}_n\|_\infty \leq 1$

D. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B : $\|\boldsymbol{x}_n\|_1 \leq 1$

E. Don't know.

Midterm question 3

A classifier has the confusion matrix given in the figure below. Which statement about the classifier is correct?



A. The Accuracy is 94% and the Error rate is 6%

B. The Accuracy is 6% and the Error rate is 94%

C. The Accuracy is 65% and the Error rate is 35%

D. There is insufficient information in the confusion matrix to determine the Accuracy and Error rate.

E. Don't know.

Midterm question 4

Which statement about crossvalidation is wrong?

A. Cross-validation can be used to estimate the generalization error.

B. Leave one out cross-validation is more computationally expensive than 10 fold crossvalidation.

C. Holding out one third of the data for validation is faster but less accurate than performing 10 fold cross-validation.

D. The same test set can be used for model selection as well as evaluation of the generalization performance of the model.

E. Don't know.

Midterm question 5

Consider a data set of four features: A, B, C, and D that are applied in a classification algorithm. The table below shows the cross-validated Error rate when using different combinations of the features.

Feature(s)	Error rate
A	0.40
B	0.45
C	0.33
D	0.42
A and B	0.20
A and C	0.25
A and D	0.34
B and C	0.29
B and D	0.42
C and D	0.40
A and B and C	0.13
A and B and D	0.17
B and C and D	0.10
A and C and D	0.15
A and B and C and D	0.28

We will apply a forward feature selection algorithm. Which feature set will the selection algorithm choose?

A. C

B. B and C and D

C. A and B

D. A and B and C

E. Don't know.

Midterm question 6

When training a decision tree we will use the classification error as impurity measure $I(t)$ given by $I(t) = 1 - \max_i[p(i|t)]$ where $p(i|t)$ denotes the fraction of data objects belonging to class i at a given node t. We will use Huntâ€™s algorithm to grow the tree and recall that the purity gain is given by:

$$\Delta = I(\text{ Parent }) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

where N is the total number of data objects at the parent node, k is the number of child nodes and $N(v_j)$ is the number of data objects associated with the child node, v_j. We will consider classification of Iris flowers into Iris-Setosa, Iris-Virginica and Iris-Versicolor. At a potential split we have:

- Before the split: 5 Iris-Setosa, 10 Iris-Virginica and 10 Iris Versicolor.

After the split

- 0 Iris-Setosa, 8 Iris-Virginica and 2 Iris-Versicolor in the left node.
- 5 Iris-Setosa, 2 Iris-Viriginica and 8 Iris-Versicolor in the right node.

Which statement is correct?

A. The purity gain is $\Delta = \frac{3}{5}$

B. The purity gain is $\Delta = \frac{3}{15}$

C. The purity gain is $\Delta = \frac{6}{25}$

D. The purity gain is $\Delta = \frac{7}{15}$

E. Don't know.

Midterm question 7

When people are well rested and take an exam their chance of passing the exam is 90%, however, when people are not well rested there chance of passing the exam is only 40%. On any given day 80% of people are well-rested. What is the chance that a person passing the test is well rested?

A. $\frac{4}{10}$

B. $\frac{8}{10}$

C. $\frac{9}{10}$

D. $\frac{10}{11}$

E. Don't know.

Midterm question 8

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values $\sigma_1 = 4$, $\sigma_2 = 2$, $\sigma_3 = 1$, and $\sigma_4 = 0$.

Which one of the following statements is wrong?

A. The first principal component accounts for more than 60% of the variation in the data.

B. The third principal component accounts for less than 5% of the variation in the data.

C. The second principal component accounts for more than 20% of the variation in the data.

D. The data can be perfectly represented in a three dimensional sub-space.

E. Don't know.

Midterm question 9

Consider the following sequence of numbers

$$x = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 3 & 4 & 4 & 5 & 14 \end{bmatrix}.$$

What is the sum of the mean, the median and the mode of these numbers, i.e. what is the value: $y = \text{mean}(x) + \text{median}(x) + \text{mode}(x)$?

A. $y = 1$

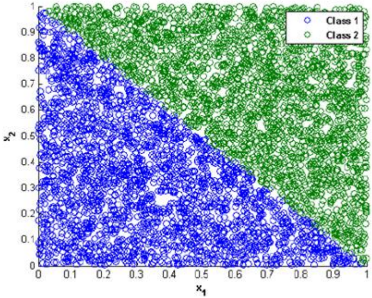
B. $y = 6$

C. $y = 7$

D. $y = 11$

E. Don't know.

Midterm question 10



Consider the classification problem given in the figure below where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w0. Which one of the following statements is **wrong**?

A. The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.

B. A decision tree with less than five nodes, all of the usual axis-aligned form $x_1 > a$ or $x_2 > b$ for different values of a, b, can perfectly separate the classes using only x_1 and x_2 as features.

C. A logistic regression model can perfectly separate the two classes using only the feature z given by $z = x_1 + x_2$.

D. In logistic regression the probability that each observation belong to the two classes can be derived from the logistic function.

E. Don't know.

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/)