Machine Learning and Intelligent Systems

Linear Models for Classification

Maria A. Zuluaga

October 20, 2023

EURECOM - Data Science Department

Table of contents

Recap

Introduction to Classification

- The Learning Process
- Parenthesis: Joint and Conditional Probabilities
- Back to Learning Intuition
- Zero-one Loss Function
- Loss Minimization
- The Bayes Classifier

Linear Discriminant Analysis

Wrap-Up

Recap

Supervised Learning: Regression

Let $y \in \mathbb{R}^O$ and $\mathbf{x} \in \mathbb{R}^D$ be related by:

$$y = f(\mathbf{x}) + \epsilon$$

We will use $O = 1$ along the course

Goal

To predict y using $\mathbf{x}$ **but** we don't know the true relationship, f, between y and $\mathbf{x}$

Linear Regression

- A linear regressor or predictor allows to find an estimate of f using x and y

Linear Regression

- A linear regressor or predictor allows to find an estimate of f using x and y
- It has the form:

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

- **w** are the **parameters** of the model

To estimate **w** we used:

- Maximum Likelihood Estimation

Linear Regression

- A linear regressor or predictor allows to find an estimate of f using x and y
- It has the form:

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

- $\mathbf{w}$ are the **parameters** of the model

To estimate $\mathbf{w}$ we used:

- Maximum Likelihood Estimation
- which leads to Ordinary Least Squares solution

Linear Regression

- A linear regressor or predictor allows to find an estimate of f using x and y
- It has the form:

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

- $\mathbf{w}$ are the **parameters** of the model

To estimate $\mathbf{w}$ we used:

- Maximum Likelihood Estimation
- which leads to Ordinary Least Squares solution

- A linear regressor or predictor allows to find an estimate of f using x and y
- It has the form:

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

- $\mathbf{w}$ are the **parameters** of the model

To estimate $\mathbf{w}$ we used:

- Maximum Likelihood Estimation
- which leads to Ordinary Least Squares solution

OLS -closed form solution:

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

Introduction to Classification

Recap: Definition

- We have $y \in \mathcal{C}$ and $\mathbf{x} \in \mathbb{R}^D$
- They are related by an unknown function $f : \mathbb{R}^D \longrightarrow \mathcal{C}$

y	$\mathbf{x}$
Output	Input
Target	Feature vector
Label	Attributes
Dependent variable	Independent variable

Goal

To predict y using $\mathbf{x}$ **but** we don't know the true relationship, f, between y and $\mathbf{x}$

Classification: Assumptions & Definitions

- **Assumption 1:** The target variable (output) y is now binary

Classification: Assumptions & Definitions

- **Assumption 1:** The target variable (output) y is now binary
  - y represents labels or classes
  - $\mathcal{C} = \{0, 1\}$, $\mathcal{C} = \{-1, +1\}$
  - $\mathcal{C} = \{0, 1, \ldots, K\}$

Classification: Assumptions & Definitions

- **Assumption 1:** The target variable (output) y is now binary
  - y represents labels or classes
  - $\mathcal{C} = \{0, 1\}$, $\mathcal{C} = \{-1, +1\}$
  - $\mathcal{C} = \{0, 1, \ldots, K\}$

- **Definition:** $\mathcal{C}$ denotes the set of possible classes that y can take
  - <u>Examples:</u>

Classification: Assumptions & Definitions

- **Assumption 1:** The target variable (output) y is now binary
  - y represents labels or classes
  - $\mathcal{C} = \{0, 1\}$, $\mathcal{C} = \{-1, +1\}$
  - $\mathcal{C} = \{0, 1, \ldots, K\}$

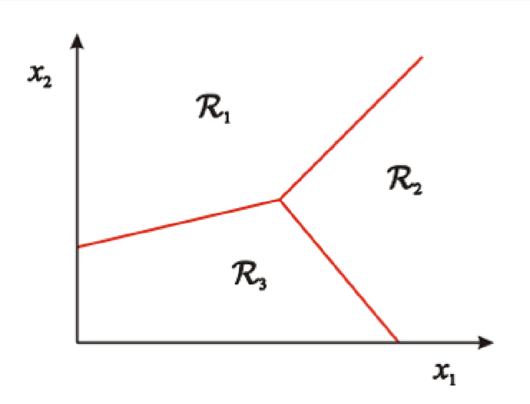
- **Definition:** $\mathcal{C}$ denotes the set of possible classes that y can take
  - Examples:

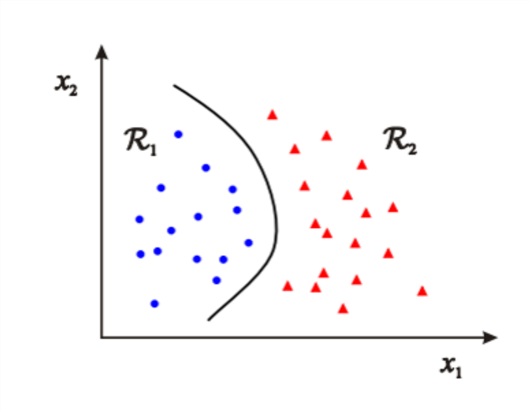
- **Assumption 2:** The input data $\mathbf{x}$ is separable

Goal

To predict the correct class $y = c \in \mathcal{C}$ using $\mathbf{x}$

Classification: Separability



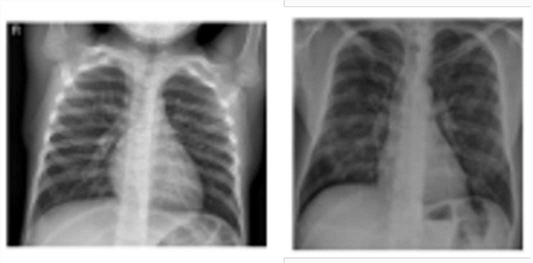


The input space divided into **decision regions** whose boundaries are called **decision boundaries** or **decision surfaces**.

An example: COVID-19 diagnosis from X-ray images

Task: To diagnose COVID-19 from X-Ray images.

- y -
- $\mathbf{x}$ -
- $\mathcal{D}$ -



Left: Healthy patient, Right: Patient with COVID-19

Source: Hammoudi et al. hal-02533605

Multi-class Classification: Representing the output variable y

- For regression problems, the target variable y was the vector of real numbers whose values we wish to predict
- In the binary case, we choose $\mathcal{C} = \{0, 1\}$ or $\mathcal{C} = \{-1, +1\}$

Multi-class Classification: Representing the output variable y

- For regression problems, the target variable y was the vector of real numbers whose values we wish to predict
- In the binary case, we choose $\mathcal{C} = \{0, 1\}$ or $\mathcal{C} = \{-1, +1\}$
- When we have multiple classes, we avoid using $\mathcal{C} = \{0, 1, \ldots, K\}$

Multi-class Classification: Representing the output variable y

- For regression problems, the target variable y was the vector of real numbers whose values we wish to predict
- In the binary case, we choose $\mathcal{C} = \{0, 1\}$ or $\mathcal{C} = \{-1, +1\}$
- When we have multiple classes, we avoid using $\mathcal{C} = \{0, 1, \dots, K\}$

- We use 1-of-K coding scheme (a.k.a one-hot encoding)

Multi-class Classification: Representing the output variable y

- For regression problems, the target variable y was the vector of real numbers whose values we wish to predict
- In the binary case, we choose $\mathcal{C} = \{0, 1\}$ or $\mathcal{C} = \{-1, +1\}$
- When we have multiple classes, we avoid using $\mathcal{C} = \{0, 1, \ldots, K\}$

- We use 1-of-K coding scheme (a.k.a one-hot encoding)
    - $\mathbf{y}$ is a vector of length K such that if the class associated to the sample $\mathbf{x}$ is the j^{th} one, then all elements y_k of $\mathbf{y}$ are zero except element y_j, which is set to 1

Multi-class Classification: Representing the output variable y

- For regression problems, the target variable y was the vector of real numbers whose values we wish to predict
- In the binary case, we choose $\mathcal{C} = \{0, 1\}$ or $\mathcal{C} = \{-1, +1\}$
- When we have multiple classes, we avoid using $\mathcal{C} = \{0, 1, \ldots, K\}$

- We use 1-of-K coding scheme (a.k.a one-hot encoding)
  - $\mathbf{y}$ is a vector of length K such that if the class associated to the sample $\mathbf{x}$ is the j^{th} one, then all elements y_k of $\mathbf{y}$ are zero except element y_j, which is set to 1
  - Example: In the slide about separability, suppose a sample $\mathbf{x}$ is assigned $\mathcal{R}_3$ as the output class. Write down the 1-of-K encoding representation:

Multi-class Classification: Representing the output variable y

- For regression problems, the target variable y was the vector of real numbers whose values we wish to predict
- In the binary case, we choose $\mathcal{C} = \{0, 1\}$ or $\mathcal{C} = \{-1, +1\}$
- When we have multiple classes, we avoid using $\mathcal{C} = \{0, 1, \ldots, K\}$

- We use 1-of-K coding scheme (a.k.a one-hot encoding)
  - $\mathbf{y}$ is a vector of length K such that if the class associated to the sample $\mathbf{x}$ is the j^{th} one, then all elements y_k of $\mathbf{y}$ are zero except element y_j, which is set to 1
  - Example: In the slide about separability, suppose a sample $\mathbf{x}$ is assigned $\mathcal{R}_3$ as the output class. Write down the 1-of-K encoding representation:

$$\mathbf{y} = (0, 0, 1)^T$$

The Learning Process: Intuition

Lets use an exercise to get an intuition of what will follow.

Problem Statement:[1]

- You are a nurse screening a set of students for a sickness calledDiseasitis.

[1] Adapted from: https://arbital.com/p/bayes_rule/?l=1zq

The Learning Process: Intuition

Lets use an exercise to get an intuition of what will follow.

Problem Statement:[1]

- You are a nurse screening a set of students for a sickness calledDiseasitis.
- We know from past studies that $\sim$20% of the students get Diseasitis at this time of year

[1] Adapted from: https://arbital.com/p/bayes_rule/?l=1zq

The Learning Process: Intuition

Lets use an exercise to get an intuition of what will follow.

Problem Statement:[1]

- You are a nurse screening a set of students for a sickness calledDiseasitis.
- We know from past studies that ∼20% of the students get Diseasitis at this time of year
- Diseasitis is tested using a color-changing tongue depressor. It turns black if the student has Diseasitis.

[1] Adapted from: `https://arbital.com/p/bayes_rule/?l=1zq`

The Learning Process: Intuition

Lets use an exercise to get an intuition of what will follow.

Problem Statement:[1]

- You are a nurse screening a set of students for a sickness calledDiseasitis.
- We know from past studies that $\sim$20% of the students get Diseasitis at this time of year
- Diseasitis is tested using a color-changing tongue depressor. It turns black if the student has Diseasitis.
- Among patients with Diseasitis, 90% turn the tongue depressor black.
- However, it also turns black 30% of the time for healthy students.

[1] Adapted from: `https://arbital.com/p/bayes_rule/?l=1zq`

The Learning Process: Intuition

Lets use an exercise to get an intuition of what will follow.

Problem Statement:[1]

- You are a nurse screening a set of students for a sickness calledDiseasitis.
- We know from past studies that ∼20% of the students get Diseasitis at this time of year
- Diseasitis is tested using a color-changing tongue depressor. It turns black if the student has Diseasitis.
- Among patients with Diseasitis, 90% turn the tongue depressor black.
- However, it also turns black 30% of the time for healthy students.
- You are tested and the tongue depressor gets black.

[1] Adapted from: `https://arbital.com/p/bayes_rule/?l=1zq`

The Learning Process: Intuition

Lets use an exercise to get an intuition of what will follow.

Problem Statement:[1]

- You are a nurse screening a set of students for a sickness calledDiseasitis.
- We know from past studies that ∼20% of the students get Diseasitis at this time of year
- Diseasitis is tested using a color-changing tongue depressor. It turns black if the student has Diseasitis.
- Among patients with Diseasitis, 90% turn the tongue depressor black.
- However, it also turns black 30% of the time for healthy students.
- You are tested and the tongue depressor gets black.
- **Question:** What is your probability of having Diseasitis?
- Hint: Your reading about Bayes rule

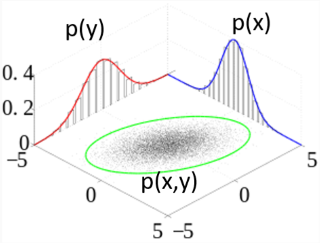
[1] Adapted from: `https://arbital.com/p/bayes_rule/?l=1zq`

Joint Probabilities

- The supervised learning problem has an input $\mathbf{X}$ and the corresponding target output vector $\mathbf{y}$ with the goal to predict y given a new value $\mathbf{x}$.
- The joint probability distribution $p(\mathbf{x}, y)$ provides a view on the uncertainty of these variables.

Joint Probabilities

- The supervised learning problem has an input $\mathbf{X}$ and the corresponding target output vector $\mathbf{y}$ with the goal to predict y given a new value $\mathbf{x}$.
- The joint probability distribution $p(\mathbf{x}, y)$ provides a view on the uncertainty of these variables.



Source: `https://en.wikipedia.org/wiki/Joint_probability_distribution`

Joint probability: For two discrete random variables, X and Y, $\mathrm{P}(X = x, Y = y)$ is the probability that random variable X has value x and random Y has value y.

Joint density function: For two continuous random variables, x and y, $p(x, y)$ is the joint density function (pdf).

Conditional Probabilities

- When variables are dependent it is possible to work with conditioning
- Example: Probability of breaking the world marathon record (B=1) given that the temperature will be above 30 (A=1)

$$P(B = 1|A = 1)$$

- Conditional PDF example

$$p(y_i|\mathbf{x}_i; \mathbf{w}, \sigma^2)$$

- Conditional probability example

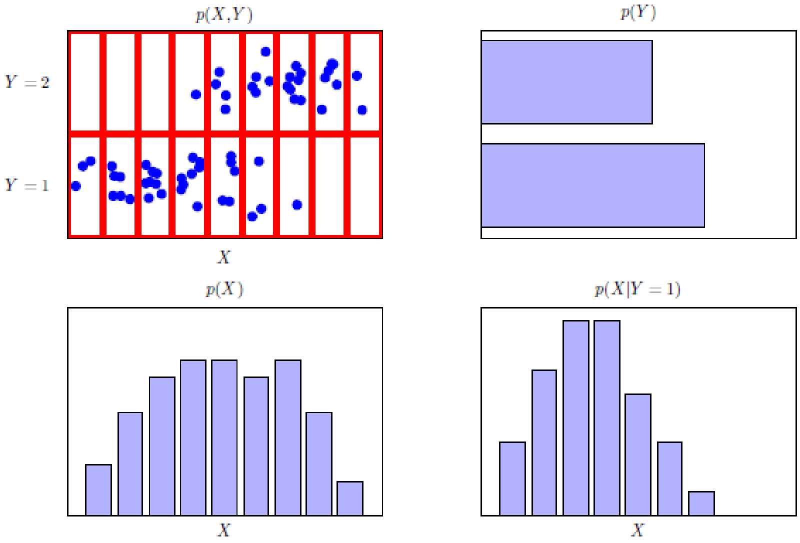
$$P(9 \leq y_i \leq 9.8|\mathbf{x}_i; \mathbf{w}, \sigma^2)$$

The Rules of Probability

	Discrete variables	**Continuous variables**
Sum rule	$P(X) = \sum_Y P(X, Y)$	$p(x) = \int_y p(x, y) dy$
Product rule	$P(X, Y) = P(Y\|X)P(X)$	$p(x, y) = p(y\|x)p(x)$

- $P(X, Y)$: Joint probability.
- $P(Y|X)$: Conditional probability, e.g. the probability of Y given X.
- $P(X)$: Marginal probability, e.g. the probability of X.

An Illustration



Source: Figure 1.11 from PRML - Bishop

Bayes' Theorem

Using the product rule and the symmetry property $P(X, Y) = P(Y, X)$ we obtain the following relationship among conditional probabilities:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}, \tag{1}$$

which is known as the **Bayes'theorem**.

Bayes' Theorem

Using the product rule and the symmetry property $P(X, Y) = P(Y, X)$ we obtain the following relationship among conditional probabilities:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}, \tag{1}$$

which is known as the **Bayes'theorem**.

Using the sum rule, the denominator in Bayes' theorem can be expressed in terms of the quantities appearing in the numerator:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_Y P(X|Y)P(Y)} \tag{2}$$

The Elements in the Bayes' theorem

Quantity	Name	Interpretation
$P(Y)$	Prior probability of Y	Probability of a hypothesis Y without any additional prior information
$P(X\|Y)$	Likelihood	Probability of observing the new evidence, given the initial hypothesis
$P(Y\|X)$	Posterior probability	Quantity of interest. Probability of Y given the evidence X
$P(X)$	Evidence or marginal likelihood	Total probability of observing the evidence

Back to the exercise

How can we express our problem in terms of the Bayes'Rule?

The Elements in the Bayes' Theorem in our Problem

Quantity	Name	Interpretation
$P(y = S)$	Prior probability of class SICK	Probability of a person having diseasitis
$P(\mathbf{x} = B \mid y = S)$	Likelihood	Probability of observing the new evidence, given initial hypothesis. Probability of the having a black tongue depressor if SICK
$p(y = S \mid \mathbf{x} = B)$	Posterior probability	Revised probability of having condition SICK after applying Bayes' theorem in light of the info contained in the tongue depressor (BLACK)
$p(\mathbf{x} = B)$	Evidence or marginal likelihood	Total probability of observing the evidence, e.g. a black tongue depressor.

The Learning Process

- Our intuition tells us that we would choose the class having the higher posterior probability to minimize the chance of assigning $\mathbf{x}$ to the wrong class.

The Learning Process

- Our intuition tells us that we would choose the class having the higher posterior probability to minimize the chance of assigning $\mathbf{x}$ to the wrong class.
- Let's formalize this.

The Learning Process

Same learning procedure as with regression:

The Learning Process

Same learning procedure as with regression:

- Collect training data $\mathcal{D}$

The Learning Process

Same learning procedure as with regression:

- Collect training data $\mathcal{D}$
- Estimate a model h using $\mathcal{D}$

The Learning Process

Same learning procedure as with regression:

- Collect training data $\mathcal{D}$
- Estimate a model h using $\mathcal{D}$
- $\hat{y} = y$ **This is different!**

The Learning Process

Same learning procedure as with regression:

- Collect training data $\mathcal{D}$
- Estimate a model h using $\mathcal{D}$
- $\hat{y} = y$ **This is different!**

We have to introduce a new loss function for the classification problem

Zero-one Loss Function

- We already know a loss function for the classification problem

Zero-one Loss Function

- We already know a loss function for the classification problem
- **Zero-one loss function:** Counts how many mistakes the estimated model h makes on the training set.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{y}_i \neq y_i}, \quad \text{where} \quad \delta_{\hat{y}_i \neq y_i} = \begin{cases} 1, & \text{if} \quad \hat{y}_i \neq y_i \\ 0, & \text{otheriwse} \end{cases} \tag{3}$$

Zero-one Loss Function

- We already know a loss function for the classification problem
- **Zero-one loss function:** Counts how many mistakes the estimated model h makes on the training set.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{y}_i \neq y_i}, \quad \text{where} \quad \delta_{\hat{y}_i \neq y_i} = \begin{cases} 1, & \text{if} \quad \hat{y}_i \neq y_i \\ 0, & \text{otheriwse} \end{cases} \tag{3}$$

- This loss function returns the error rate of the data set $\mathcal{D}$.

The Zero-one Loss Function

The 0-1 Loss Function can be seen as a $L : K \times K$ matrix, with $K = |\mathcal{C}|$.

Example: In our COVID-19 example:

The Zero-one Loss Function

The 0-1 Loss Function can be seen as a $L : K \times K$ matrix, with $K = |\mathcal{C}|$.

Example: In our COVID-19 example:

	COVID-19	Healthy
COVID-19		
Healthy		

The Zero-one Loss Function

The 0-1 Loss Function can be seen as a $L : K \times K$ matrix, with $K = |\mathcal{C}|$.

Example: In our COVID-19 example:

	COVID-19	Healthy
COVID-19		
Healthy		

The 0-1 loss function might not always be the best choice. Do you see any disadvantages of using it in this example?

Loss Minimization

- Our goal is to minimize the average loss
- We can use the definition of expectation to express the average loss:

$$\mathbb{E}[x] = \sum_{X} XP(X)$$

Loss Minimization

- Our goal is to minimize the average loss
- We can use the definition of expectation to express the average loss:

$$\mathbb{E}[x] = \sum_{X} XP(X)$$

- Who is our X in this case?

Loss Minimization

- Our goal is to minimize the average loss
- We can use the definition of expectation to express the average loss:

$$\mathbb{E}[x] = \sum_X XP(X)$$

- Who is our X in this case? $\Rightarrow \quad L : K \times K$ matrix

Loss Minimization

- Our goal is to minimize the average loss
- We can use the definition of expectation to express the average loss:

$$\mathbb{E}[x] = \sum_X XP(X)$$

- Who is our X in this case? $\Rightarrow \quad L : K \times K$ matrix
- Since we are representing L as a matrix, the sum term needs to go over all elements in L.

Loss Minimization

- Our goal is to minimize the average loss
- We can use the definition of expectation to express the average loss:

$$\mathbb{E}[x] = \sum_X XP(X)$$

- Who is our X in this case? $\Rightarrow \quad L : K \times K$ matrix
- Since we are representing L as a matrix, the sum term needs to go over all elements in L.
- We will denote L_{kj}, with k the index of the true class and j the class to which $\mathbf{x}$ is being assigned to (which may be equal to k or not)

$$\sum_k \sum_j L_{kj}$$

Loss Minimization (2)

- The loss depends on the true class, which is unknown.

Loss Minimization (2)

- The loss depends on the true class, which is unknown.
- For a given $\mathbf{x}$, the uncertainty in the true class is expressed through the joint probability distribution $p(\mathbf{x}, \mathcal{C}_k)$.

Loss Minimization (2)

- The loss depends on the true class, which is unknown.
- For a given $\mathbf{x}$, the uncertainty in the true class is expressed through the joint probability distribution $p(\mathbf{x}, \mathcal{C}_k)$.
- So, the average loss is computed with respect to this distribution.
- Putting all terms together, we get the expected loss:

$$\mathbb{E}[\mathcal{L}] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}$$

Loss Minimization (2)

- The loss depends on the true class, which is unknown.
- For a given $\mathbf{x}$, the uncertainty in the true class is expressed through the joint probability distribution $p(\mathbf{x}, \mathcal{C}_k)$.
- So, the average loss is computed with respect to this distribution.
- Putting all terms together, we get the expected loss:

$$\mathbb{E}[\mathcal{L}] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}$$

- The goal is to choose an $\mathcal{R}_j$ that minimizes the expected loss.

Loss Minimization (3)

- The $\mathcal{R}_j$ to which $\mathbf{x}$ is assigned should minimize

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$$

Loss Minimization (3)

- The $\mathcal{R}_j$ to which $\mathbf{x}$ is assigned should minimize

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$$

- Refactoring with the product rule, $p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$, the decision rule that minimizes the expected loss is one assigning $\mathbf{x}$ to class j for which

$$\min \sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$$

An Example with the 0/1 Loss

- The expression we have obtained is generic ($\mathcal{L}$) and not attached to the zero-one loss.
- It allows to change the penalization associated to each type of error.
- COVID-19 example: Where is worse to make a mistake?

An Example with the 0/1 Loss

- The expression we have obtained is generic ($\mathcal{L}$) and not attached to the zero-one loss.
- It allows to change the penalization associated to each type of error.
- COVID-19 example: Where is worse to make a mistake?

- Let's have a look at the minimization problem using the 0-1 loss
- For K classes, and k the index of the correct class we have:

$$L_{k0}\,p(\mathcal{C}_0|\mathbf{x}) + L_{k1}\,p(\mathcal{C}_1|\mathbf{x}) + \ldots + L_{kk}\,p(\mathcal{C}_k|\mathbf{x}) + \ldots + L_{kK}\,p(\mathcal{C}_K|\mathbf{x})$$

An Example with the 0/1 Loss (cont)

Since k is the correct class, the term associated to it cancels out:

$$L_{k0}p(\mathcal{C}_0|\mathbf{x}) + L_{k1}p(\mathcal{C}_1|\mathbf{x}) + \ldots + \ldots + L_{kK}p(\mathcal{C}_K|\mathbf{x}) = \sum_{j \neq k} p(\mathcal{C}_j|\mathbf{x})$$

An Example with the 0/1 Loss (cont)

Since k is the correct class, the term associated to it cancels out:

$$L_{k0}p(\mathcal{C}_0|\mathbf{x}) + L_{k1}p(\mathcal{C}_1|\mathbf{x}) + \ldots + \ldots + L_{kK}p(\mathcal{C}_K|\mathbf{x}) = \sum_{j \neq k} p(\mathcal{C}_j|\mathbf{x})$$

and the sum of all conditional probabilities is:

$$\sum_{j}^{K} p(\mathcal{C}_j|\mathbf{x}) = 1$$

An Example with the 0/1 Loss (cont)

Since k is the correct class, the term associated to it cancels out:

$$L_{k0}p(\mathcal{C}_0|\mathbf{x}) + L_{k1}p(\mathcal{C}_1|\mathbf{x}) + \ldots + \ldots + L_{kK}p(\mathcal{C}_K|\mathbf{x}) = \sum_{j \neq k} p(\mathcal{C}_j|\mathbf{x})$$

and the sum of all conditional probabilities is:

$$\sum_{j}^{K} p(\mathcal{C}_j|\mathbf{x}) = 1$$

the minimization problem (for the 0-1 loss) can be expressed as:

$$\hat{y} = \arg\min_{j} 1 - p(\mathcal{C}_j|\mathbf{x})$$

An Example with the 0/1 Loss (cont)

Since k is the correct class, the term associated to it cancels out:

$$L_{k0}p(\mathcal{C}_0|\mathbf{x}) + L_{k1}p(\mathcal{C}_1|\mathbf{x}) + \ldots + \ldots + L_{kK}p(\mathcal{C}_K|\mathbf{x}) = \sum_{j \neq k} p(\mathcal{C}_j|\mathbf{x})$$

and the sum of all conditional probabilities is:

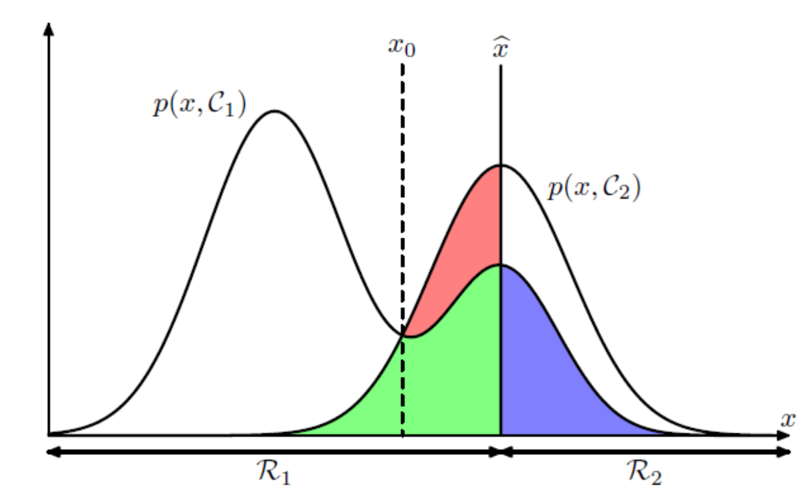
$$\sum_{j}^{K} p(\mathcal{C}_j|\mathbf{x}) = 1$$

the minimization problem (for the 0-1 loss) can be expressed as:

$$\begin{aligned} \hat{y} &= \arg\min_j 1 - p(\mathcal{C}_j|\mathbf{x}) \\ &= \arg\max_j p(\mathcal{C}_j|\mathbf{x}) \end{aligned}$$

Bayes classifier: We classify to the most likely class using the conditional distribution

Illustration: Two Classes



Source: Figure 1.24 from PRML - Bishop

The Bayes Classifier

- We classify to the most likely class using the conditional distribution

$$
\begin{aligned}
\hat{y} &= \arg\min_j 1 - p(\mathcal{C}_j|\mathbf{x}) \\
&= \arg\max_j p(\mathcal{C}_j|\mathbf{x})
\end{aligned}
$$

The Bayes Classifier

- We classify to the most likely class using the conditional distribution

$$\hat{y} = \arg\min_j 1 - p(\mathcal{C}_j|\mathbf{x})$$
$$= \arg\max_j p(\mathcal{C}_j|\mathbf{x})$$

- as long as we know $p(\mathcal{C}_j|\mathbf{x})$

The Bayes Classifier

- We classify to the most likely class using the conditional distribution

$$
\begin{aligned}
\hat{y} &= \arg\min_j 1 - p(\mathcal{C}_j|\mathbf{x}) \\
&= \arg\max_j p(\mathcal{C}_j|\mathbf{x})
\end{aligned}
$$

- as long as we know $p(\mathcal{C}_j|\mathbf{x})$

- We will now introduce two different (linear) methods to estimate $p(\mathcal{C}_j|\mathbf{x})$:
  - Linear Discriminant Analysis
  - Logistic Regression

Linear Discriminant Analysis

According to the Bayes classifier we need to have an estimate of $p(\mathcal{C}_k|\mathbf{x})$ to be able to use it.

Introduction

According to the Bayes classifier we need to have an estimate of $p(\mathcal{C}_k|\mathbf{x})$ to be able to use it. We are going to use the Bayes' theorem to estimate $p(\mathcal{C}_k|\mathbf{x})$:

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

under certain assumptions about the data and the model.

Data & Model Assumptions

Data Assumptions:

1. Data within each class is normally distributed:

$$p(\mathbf{x}|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma}) = f_k(\mathbf{x})$$

Data & Model Assumptions

Data Assumptions:

1. Data within each class is normally distributed:

$$p(\mathbf{x}|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma}) = f_k(\mathbf{x})$$

2. Each class has its own mean μ_k, but they all share a common covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{D \times D}$

$$f_k(\mathbf{x}) = \frac{1}{2\pi^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x} - \mu_k\right)^T \mathbf{\Sigma}^{-1}\left(\mathbf{x} - \mu_k\right)\right)$$

Data & Model Assumptions

Data Assumptions:

1. Data within each class is normally distributed:

$$p(\mathbf{x}|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma}) = f_k(\mathbf{x})$$

2. Each class has its own mean μ_k, but they all share a common covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{D \times D}$

$$f_k(\mathbf{x}) = \frac{1}{2\pi^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu_k)\right)$$

Model Assumptions:

Linear model in $\mathbf{x}$

Formulation

- The Bayes classifier states: $\hat{y} = \arg\max_{k} p(\mathcal{C}_k|\mathbf{x})$

Formulation

- The Bayes classifier states: $\hat{y} = \arg\max_k p(\mathcal{C}_k|\mathbf{x})$
- The Bayes'theorem: $p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{f_k(\mathbf{x})p(\mathcal{C}_k)}{p(\mathbf{x})}$

Formulation

- The Bayes classifier states: $\hat{y} = \arg\max_{k} p(\mathcal{C}_k|\mathbf{x})$
- The Bayes'theorem: $p(\mathcal{C}_k|\mathbf{x}) = \dfrac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \dfrac{f_k(\mathbf{x})p(\mathcal{C}_k)}{p(\mathbf{x})}$
- Since the denominator is common to all classes, we can suppress it and define the following decision rule:

$$\hat{y}_{LDA} = \arg\max_{k \in \mathcal{C}} f_k(\mathbf{x})p(\mathcal{C}_k)$$

Formulation

- The Bayes classifier states: $\hat{y} = \arg\max_{k} p(\mathcal{C}_k|\mathbf{x})$
- The Bayes'theorem: $p(\mathcal{C}_k|\mathbf{x}) = \dfrac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \dfrac{f_k(\mathbf{x})p(\mathcal{C}_k)}{p(\mathbf{x})}$
- Since the denominator is common to all classes, we can suppress it and define the following decision rule:

$$\hat{y}_{LDA} = \arg\max_{k \in \mathcal{C}} f_k(\mathbf{x})p(\mathcal{C}_k)$$

- Since the $\log(\cdot)$ is a monotone function, we can equivalently maximize $\log(f_k(\mathbf{x})p(\mathcal{C}_k))$

Formulation

- The Bayes classifier states: $\hat{y} = \arg\max_{k} p(\mathcal{C}_k|\mathbf{x})$
- The Bayes'theorem: $p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{f_k(\mathbf{x})p(\mathcal{C}_k)}{p(\mathbf{x})}$
- Since the denominator is common to all classes, we can suppress it and define the following decision rule:

$$\hat{y}_{LDA} = \arg\max_{k \in \mathcal{C}} f_k(\mathbf{x})p(\mathcal{C}_k)$$

- Since the $\log(\cdot)$ is a monotone function, we can equivalently maximize $\log(f_k(\mathbf{x})p(\mathcal{C}_k))$
- We define:

$$\delta_k(\mathbf{x}) = \log(f_k(\mathbf{x})\pi_k)$$

Linear discriminant function

Note: We use $\pi_k = p(\mathcal{C}_k)$ to be consistent with the literature.

Derivation

$$\delta_k(\mathbf{x}) = \log(f_k(\mathbf{x})\pi_k) = \log(f_k(\mathbf{x})) + \log(\pi_k)$$

Derivation

$$\delta_k(\mathbf{x}) = \log(f_k(\mathbf{x})\pi_k) = \log(f_k(\mathbf{x})) + \log(\pi_k)$$

$$\delta_k(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\boldsymbol{\Sigma}^{-1}\mu_k + \log(\pi_k)$$

Linear in x

Implementation

- The parameters of the Gaussian distribution and the class priors π_k are not known.
- They are estimated using the training data $\mathcal{D}$, $|\mathcal{D}| = N$:

$$\hat{\pi}_k = \frac{N_k}{N} \quad \text{- Proportion of observations of class } k$$

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{y_i \in k} \mathbf{x}_i \quad \text{- Centroids of class } k$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{y_i = k} (\mathbf{x}_i - \hat{\mu}_k)^T (\mathbf{x}_i - \hat{\mu}_k) \quad \text{- Pooled sampled covariance matrix}$$

Implementation

- The parameters of the Gaussian distribution and the class priors π_k are not known.
- They are estimated using the training data $\mathcal{D}$, $|\mathcal{D}| = N$:

$$\hat{\pi}_k = \frac{N_k}{N} \quad \text{- Proportion of observations of class } k$$

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{y_i \in k} \mathbf{x}_i \quad \text{- Centroids of class } k$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{y_i = k} (\mathbf{x}_i - \hat{\mu}_k)^T (\mathbf{x}_i - \hat{\mu}_k) \quad \text{- Pooled sampled covariance matrix}$$

- A prediction is obtained by replacing:

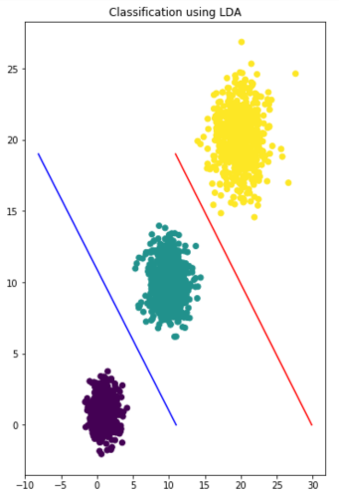
$$\hat{\delta}_k(\mathbf{x}) = \mathbf{x}^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k + \log(\hat{\pi}_k)$$

$$\hat{y}_{LDA} = \underset{k \in \mathcal{C}}{\arg\max}\, \hat{\delta}_k(\mathbf{x})$$

Decision Boundary

The decision boundary between two classes j, k is found where:

$$\delta_k(\mathbf{x}) = \delta_j(\mathbf{x}) \qquad (4)$$



See: `02_linear_classifiers.ipynb`

Decision Boundary

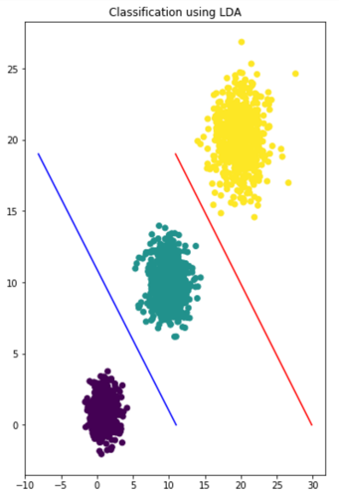
The decision boundary between two classes j, k is found where:

$$\delta_k(\mathbf{x}) = \delta_j(\mathbf{x}) \tag{4}$$

From the linear discriminant function:

$$\hat{\delta}_k(\mathbf{x}) = \mathbf{x}^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k^T + \log(\hat{\pi}_k)$$

one can see it follows the form $a_k + b_k^T x$.



See: `02_linear_classifiers.ipynb`

Decision Boundary

The decision boundary between two classes j, k is found where:

$$\delta_k(\mathbf{x}) = \delta_j(\mathbf{x}) \tag{4}$$

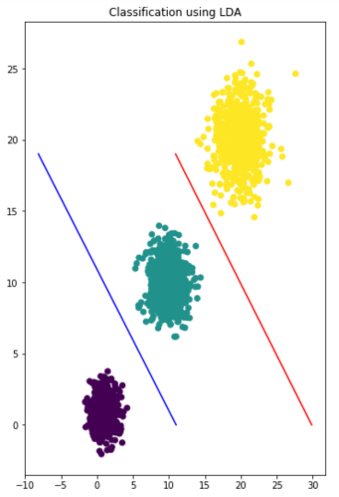
From the linear discriminant function:

$$\hat{\delta}_k(\mathbf{x}) = \mathbf{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\mu}_k^T + \log(\hat{\pi}_k)$$

one can see it follows the form $a_k + b_k^T x$.
Replacing in Eq. 4:

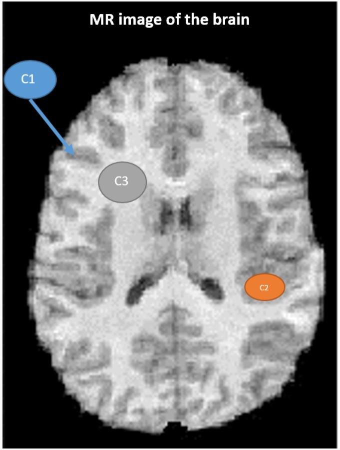
$$a_k + b_k^T x = a_j + b_j^T x$$
$$(a_k - a_j) + (b_k^T - b_j^T)x = 0$$

Exercise: Obtain generic expressions for the decision boundaries in the figure, where $K = 3$ and $\mathbf{x} \in \mathbb{R}^2$.



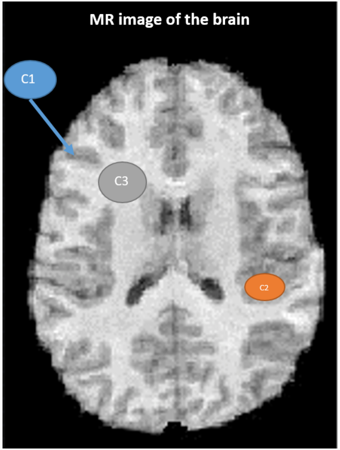
See: `02_linear_classifiers.ipynb`

Exercise: Brain Segmentation



The brain is composed of gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF).

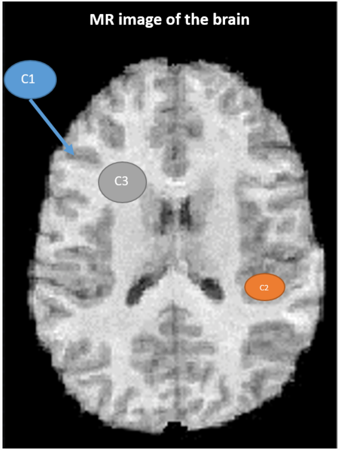
Exercise: Brain Segmentation



The brain is composed of gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF).
The 3 have different appearances when imaged using magnetic resonance (MR) imaging:

- C1: CSF, C2: GM and C3: WM

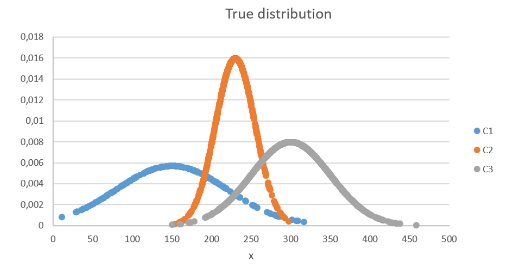
Exercise: Brain Segmentation



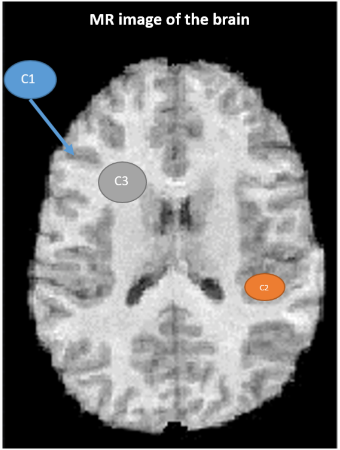
The brain is composed of gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF).
The 3 have different appearances when imaged using magnetic resonance (MR) imaging:

- C1: CSF, C2: GM and C3: WM

The true joint distributions of the image intensities are shown below:



Exercise: Brain Segmentation



Your tasks:

- Use LDA to estimate the posteriors $p(\mathcal{C}_1|x)$, $p(\mathcal{C}_2|x)$ and $p(\mathcal{C}_3|x)$.
- Compare the obtained results with the true distributions. What can you say about LDA?
- Use `lda_playground.xls` to estimate the parameters and implement the necessary functions.
- The training data is provided in the file.

Wrap-Up

Recap

In this lecture...

- We introduced the concept of classification
- We reviewed some useful concepts from probability which can be applied to decision theory
- We introduced the learning process in classification problems
- We saw a first learning algorithm for classification: Linear Discriminant Analysis (LDA)

Key Concepts

- Discrete output, target
- 1-of-K encoding
- Joint probability
- Bayes' theorem
- Classification
- Zero-one loss function
- Linear Discriminant Analysis (LDA)

References

Further Reading and Useful Material

Source	Notes
Pattern Recognition and Machine Learning	Ch. 4
The Elements of Statistical Learning	Ch. 2 and 4