

# \*\*\* Applied Machine Learning Fundamentals \*\*\*

## Decision Theory

Daniel Wehner

SAP SE

August 22, 2019



# Agenda August 22, 2019

## ① Bayesian Decision Theory

- Introduction
- Class Conditional Probabilities
- Class Priors
- Bayes' Theorem
- Bayes' Optimal Classifier

## ② Naïve Bayes Classifier

- Assumptions and Algorithm

An Example

Laplace Smoothing

## ③ Risk Minimization

## ④ Wrap-Up

Summary

Lecture Overview

Self-Test Questions

Recommended Literature and further Reading

Section:  
**Bayesian Decision Theory**

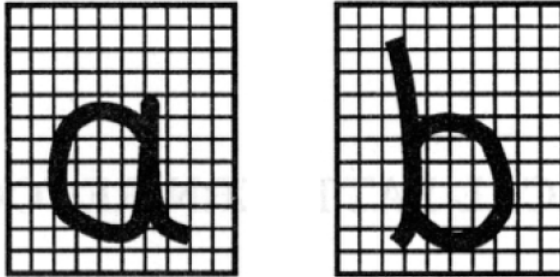


# Statistical Methods

- Statistical methods assume that the process that 'generates' the data is governed by the **rules of probability**
- The data is understood to be a set of **random samples** from some underlying **probability distribution**
- This is the reason for the name **statistical machine learning**

The basic assumption about how the data is generated is always there, even if you don't see a single probability distribution!

# Running Example: Optical Character Recognition (OCR)



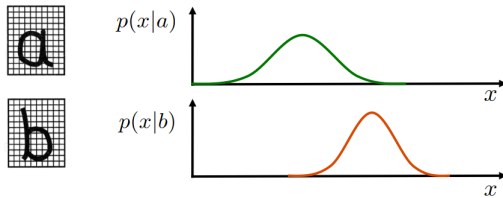
**Goal: Classify a new letter so that the probability of a wrong classification is minimized**

# Class Conditional Probabilities

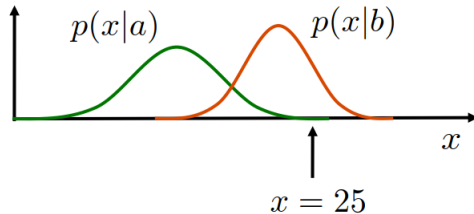
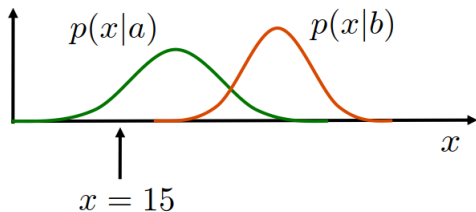
- First concept: **Class conditional probabilities**
- Probability of  $\mathbf{x}$  given a specific class  $\mathcal{C}_k$  is formally written as:

$$p(\mathbf{x}|\mathcal{C}_k) \in [0, 1] \quad (1)$$

- $\mathbf{x} \in \mathbb{R}^m$  is a feature vector, e. g. # black pixels, height-width ratio, ...



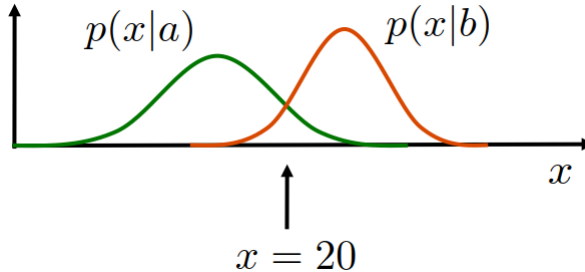
## Class Conditional Probabilities (Ctd.)



If  $x = 15$  we would predict class  $a$  since  $p(15|a) > p(15|b)$ .

If  $x = 25$  we would output class  $b$  since  $p(25|b) > p(25|a)$ .

## Class Conditional Probabilities (Ctd.)



We have a problem!

- Which class should be chosen now?
- The conditional probabilities are the same... ☠



# Class Prior Probabilities

- Second concept: **Class priors**
- The prior probability of a data point belonging to a particular class  $\mathcal{C}$

$$\mathcal{C}_1 \equiv a \quad p(\mathcal{C}_1) = 0.75$$

$$\mathcal{C}_2 \equiv b \quad p(\mathcal{C}_2) = 0.25$$

- By definition:

How would you decide now?

- $0 \leq p(\mathcal{C}_k) \leq 1, \forall k$
  - The sum of all probabilities equals one:  $\sum_{k=1}^{|\mathcal{C}|} p(\mathcal{C}_k) = 1$
- **The class prior is equivalent to a prior belief in the class label**

# How to get the Prior Probabilities?

Count Count's advice:

Simply count the  
number of instances  
in each class!

But don't count apples!



# Bayes' Theorem

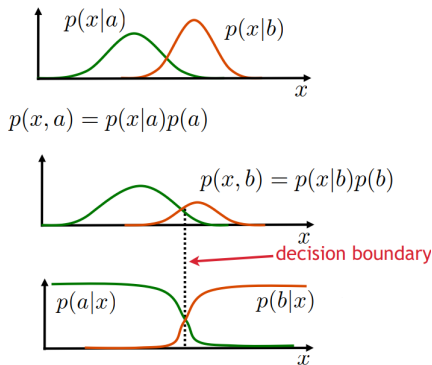
- What we actually want to compute:  $P(\mathcal{C}_k|\mathbf{x}) \Rightarrow$  **Posterior probability**
- We can compute it by applying **Bayes' theorem**
- This is one of the **most important formulas (!!!)**

$$\begin{array}{c} \text{Class posterior} \\ \overbrace{p(\mathcal{C}_k|\mathbf{x})} \end{array} = \frac{\overbrace{p(\mathbf{x}|\mathcal{C}_k)}^{\text{Class cond.}} \cdot \overbrace{p(\mathcal{C}_k)}^{\text{Class prior}}}{\underbrace{p(\mathbf{x})}_{\text{Normalization term}}} = \frac{p(\mathbf{x}|\mathcal{C}_k) \cdot p(\mathcal{C}_k)}{\sum_{j=1}^{|\mathcal{C}|} p(\mathbf{x}|\mathcal{C}_j) \cdot p(\mathcal{C}_j)} \quad (2)$$

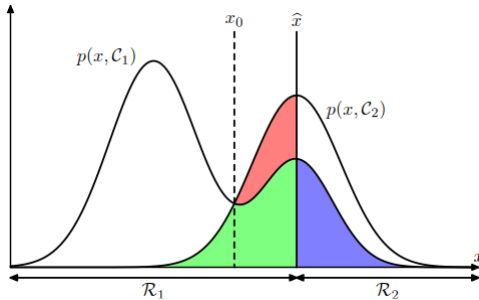
# Calculation of the Posterior Probability

- By applying Bayes' theorem we can compute the posterior
- Simply plug ❶ and ❷ into Bayes' theorem
  - Class prior probabilities
  - Class conditional probabilities

We get the final **decision boundary**



# Error Minimization



$$p(\text{error}) = p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1)$$

$$= \overbrace{\int_{\mathcal{R}_1} p(x|\mathcal{C}_2) \cdot p(\mathcal{C}_2) dx}^{\text{red + green area}} + \underbrace{\int_{\mathcal{R}_2} p(x|\mathcal{C}_1) \cdot p(\mathcal{C}_1) dx}_{\text{blue area}}$$

# Bayes' Optimal Classifier

- Decision rule:
  - Decide  $\mathcal{C}_1$  if  $p(\mathcal{C}_1|\mathbf{x}) > p(\mathcal{C}_2|\mathbf{x})$
  - This is equivalent to: *(we don't need the normalization)*

$$p(\mathbf{x}|\mathcal{C}_1) \cdot p(\mathcal{C}_1) > p(\mathbf{x}|\mathcal{C}_2) \cdot p(\mathcal{C}_2) \quad (3)$$

- Which is in turn equivalent to:

$$\frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} > \frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)} \quad (4)$$

- A classifier obeying this rule is called **Bayes' Optimal Classifier**

Section:  
**Naïve Bayes Classifier**



# A naïve Assumption

- We want to compute  $p(\mathcal{C}_k|\mathbf{x})$ . Recall Bayes' theorem:

Our first classification algorithm!

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{C}_k) \cdot P(\mathcal{C}_k)}{P(\mathbf{x})} \quad (5)$$

- Assumptions:
  - All  $x_i \in \mathbf{x}$  are **pairwise conditionally independent** ( $\Rightarrow$  naïve)

$$p(\mathbf{x}|\mathcal{C}_k) = p(x_1|\mathcal{C}_k) \cdot p(x_2|\mathcal{C}_k, x_1) \cdot p(x_3|\mathcal{C}_k, x_1, x_2) \cdot \dots = \prod_{j=1}^m p(x_j|\mathcal{C}_k) \quad (6)$$

- $p(\mathbf{x})$  is constant w. r. t. class label  $\Rightarrow$  **It is omitted**



# How to get the most probable Class?

- **Given:**
  - New instance  $\mathbf{x} = \langle x_1, x_2, \dots, x_m \rangle$  to be classified
  - Finite set of  $\ell$  classes  $\{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_\ell\}$
  - **Labeled** training data ( $\Rightarrow$  supervised learning)
- **Wanted:** Most probable class  $\mathcal{C}_{MAP}$  (maximum a posteriori) for  $\mathbf{x}$ :

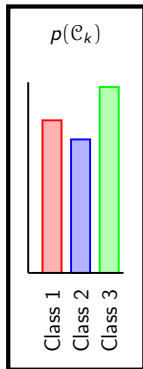
$$\mathcal{C}_{MAP} = \arg \max_{\mathcal{C}_k \in \{\mathcal{C}_1, \dots, \mathcal{C}_\ell\}} \hat{p}(\mathcal{C}_k | \mathbf{x}) \quad (7)$$

$\hat{p}$  denotes an  
**approximated** probability

$$= \arg \max_{\mathcal{C}_k \in \{\mathcal{C}_1, \dots, \mathcal{C}_\ell\}} \hat{p}(\mathcal{C}_k) \prod_{j=1}^m \hat{p}(x_j | \mathcal{C}_k) \quad (8)$$

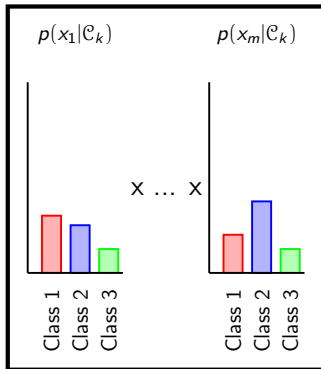
# How to get the most probable Class? (Ctd.)

Apriori Probabilities



x

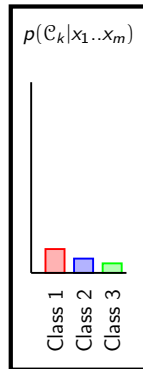
Feature Contributions



x ... x

Aposteriori Probabilities

=



# Example Data Set

Outlook	Temperature	Humidity	Wind	PlayGolf
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rainy	mild	high	weak	yes
rainy	cool	normal	weak	yes
rainy	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rainy	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rainy	mild	high	strong	no
sunny	cool	high	strong	???

# How to estimate the Probabilities?

- How to estimate the probabilities  $\hat{p}(\mathcal{C}_k)$  and  $\hat{p}(x_j|\mathcal{C}_k)$  ?
- **Solution:** Simply count the occurrences



$$\hat{p}(\mathcal{C}_k) = \frac{\sum_{i=1}^n \mathbb{1}\{y^{(i)} = \mathcal{C}_k\}}{n} \quad (9)$$

$$\hat{p}(x_j = v|\mathcal{C}_k) = \frac{\sum_{i=1}^n \mathbb{1}\{x_j^{(i)} = v \wedge y^{(i)} = \mathcal{C}_k\}}{\sum_{i=1}^n \mathbb{1}\{y^{(i)} = \mathcal{C}_k\}} \quad (10)$$

- $\mathbb{1}\{bool\}$  is the **indicator function**  
 (returns 1 if *bool* is true, 0 otherwise. E. g.:  $\mathbb{1}\{1 + 1 = 2\} = 1$ ,  $\mathbb{1}\{3 = 2\} = 0$ )

## Let's compute some Probabilities

- New instance  $\mathbf{x} = \langle \text{sunny}, \text{cool}, \text{high}, \text{strong} \rangle$
- What is its class?
- Let's compute some of the probabilities needed:

$$\hat{p}(\text{Golf} = \text{yes}) = 9/14 = 0.64$$

$$\hat{p}(\text{Golf} = \text{no}) = 5/14 = 0.36$$

$$\hat{p}(\text{Outlook} = \text{sunny} | \text{Golf} = \text{yes}) = 2/9 = 0.22$$

$$\hat{p}(\text{Outlook} = \text{sunny} | \text{Golf} = \text{no}) = 3/5 = 0.60$$

...

# Class Prediction

$$\begin{aligned}\hat{p}(\text{yes}|\mathbf{x}) &= \overbrace{\hat{p}(\text{sunny}|\text{yes})}^{=0.22} \cdot \overbrace{\hat{p}(\text{cool}|\text{yes}) \cdot \hat{p}(\text{high}|\text{yes}) \cdot \hat{p}(\text{strong}|\text{yes})}^{\text{calculate probabilities accordingly}} \cdot \overbrace{\hat{p}(\text{yes})}^{=0.64} \\ &= \mathbf{0.0053}\end{aligned}$$

$$\begin{aligned}\hat{p}(\text{no}|\mathbf{x}) &= \overbrace{\hat{p}(\text{sunny}|\text{no})}^{=0.60} \cdot \overbrace{\hat{p}(\text{cool}|\text{no}) \cdot \hat{p}(\text{high}|\text{no}) \cdot \hat{p}(\text{strong}|\text{no})}^{\text{calculate probabilities accordingly}} \cdot \overbrace{\hat{p}(\text{no})}^{=0.36} \\ &= \mathbf{0.0206}\end{aligned}$$

**Classification:**  $\mathcal{C}_{MAP} = \text{no}$  (No golf today...)

# Scaling the Output

- **But wait!** These probabilities don't sum up to one!?!?
- This is because we dropped the normalization term  $p(\mathbf{x})$
- **Scaling** can fix this:

$$\hat{p}(\text{yes}|\mathbf{x})_{\text{norm}} = \frac{0.0053}{0.0053 + 0.0206} = \mathbf{0.205}$$

$$\hat{p}(\text{no}|\mathbf{x})_{\text{norm}} = \frac{0.0206}{0.0053 + 0.0206} = \mathbf{0.795}$$

- Scaling does **not** change the prediction

# Laplace Smoothing

- **Problem:** A feature value  $v^*$  in the test data not seen during training
- $\hat{p}(v^*|\mathcal{C}_k) = 0$ : The whole product becomes zero...
- **Solution:** Laplace smoothing

$$\hat{p}(\mathcal{C}_k) = \frac{\sum_{i=1}^n \mathbb{1}\{y^{(i)} = \mathcal{C}_k\} + 1}{n + \ell} \quad (11)$$

$$\hat{p}(x_j = v|\mathcal{C}_k) = \frac{\sum_{i=1}^n \mathbb{1}\{x_j^{(i)} = v \wedge y^{(i)} = \mathcal{C}_k\} + 1}{\sum_{i=1}^n \mathbb{1}\{y^{(i)} = \mathcal{C}_k\} + \ell} \quad (12)$$



Section:  
**Risk Minimization**



# Error $\neq$ Risk

- So far, we have tried to minimize the misclassification rate
- Nevertheless, there are cases where not every misclassification is equally bad
- Some classical examples:
  - **Smoke detector**
    - If there is a fire, we must make sure to detect it
    - If there is not, an occasional false alarm may be acceptable
  - **Medical diagnosis**
    - If the patient is sick, we have to detect the disease
    - If they are healthy, it can be okay to classify them as sick (order further tests)

Section:  
**Wrap-Up**



# Summary

# Lecture Overview

## Unit I: Machine Learning Introduction

# Self-Test Questions

# Recommended Literature and further Reading

# Thank you very much for the attention!

**Topic:** \*\*\* Applied Machine Learning Fundamentals \*\*\* Decision Theory

**Date:** August 22, 2019

**Contact:**

Daniel Wehner (D062271)

SAP SE

[daniel.wehner@sap.com](mailto:daniel.wehner@sap.com)

## Do you have any questions?