*** Applied Machine Learning Fundamentals ***
Decision Theory

Daniel Wehner

SAP SE

August 22, 2019

Agenda August 22, 2019

Section:

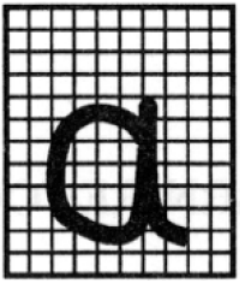
Bayesian Decision Theory

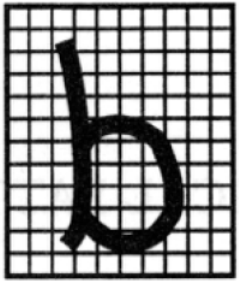
Statistical Methods

- 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)





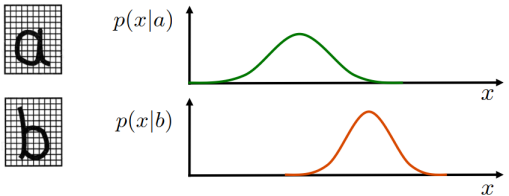
Goal: Classify a new letter so that the probability of a wrong classification is minimized

Class Conditional Probabilities

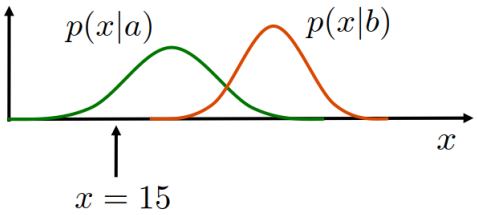
- First concept: **Class conditional probabilities**
- Probability of $\boldsymbol{x}$ given a specific class $\mathcal{C}_k$ is formally written as:

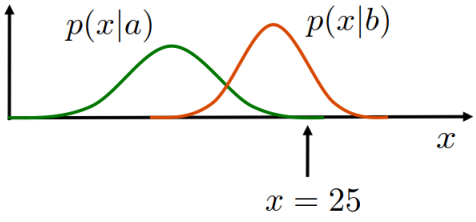
$$p(\boldsymbol{x}|\mathcal{C}_k) \in [0, 1] \tag{1}$$

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



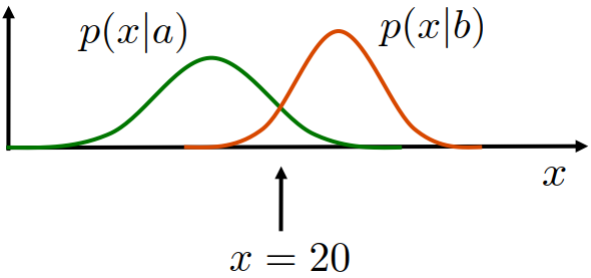
Class Conditional Probabilities (Ctd.)





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 $\mathcal{C}$

$$\mathcal{C}_1 \equiv a \qquad p(\mathcal{C}_1) = 0.75$$
$$\mathcal{C}_2 \equiv b \qquad p(\mathcal{C}_2) = 0.25$$

How would you decide now?

- By definition:
  - $0 \leqslant p(\mathcal{C}_k) \leqslant 1, \ \forall k$
  - 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!

Bayesian Decision Theory: Bayes' Theorem

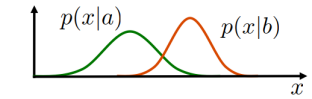
- What we actually want to compute: $P(\mathcal{C}_k|\boldsymbol{x}) \Rightarrow$ **Posterior probability**
- We can compute it by applying **Bayes' theorem**
- This is one of the **most important formulas (!!!)**

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

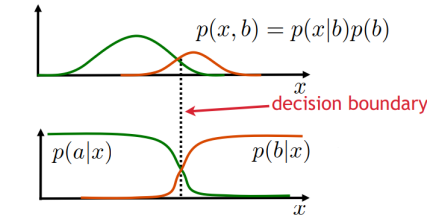
Calculation of the Posterior Probability

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

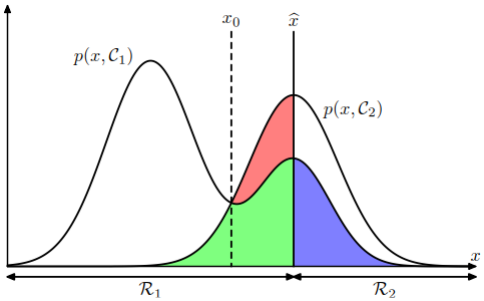
We get the final **decision boundary**



$$p(x, a) = p(x|a)p(a)$$



Bayesian Decision Theory: Bayes' Optimal Classifier (Ctd.)



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

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

Bayesian Decision Theory: Bayes' Optimal Classifier (Ctd.)

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

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

  - Which is in turn equivalent to:

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

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

Section:

Naïve Bayes Classifier

A naïve assumption

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

Our first classification algorithm!

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

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

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

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

How to get the most probable Class?

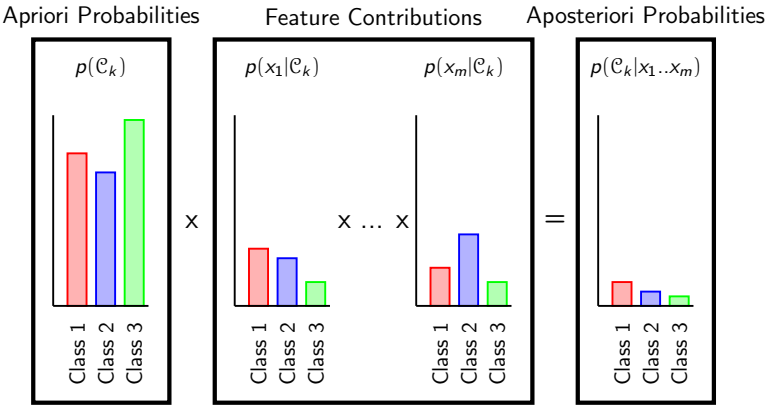
- **Given:**
  - New instance $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_m \rangle$ to be classified
  - Finite set of classes $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_l\}$
  - Labeled training data ($\Rightarrow$ supervised learning)
- **Wanted:** Most probable class $\mathcal{C}_{MAP}$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

$$\mathcal{C}_{MAP} = \arg\max_{\mathcal{C}_k \in \mathcal{C}} \widehat{p}(\mathcal{C}_k|\boldsymbol{x}) \tag{7}$$

$$= \arg\max_{\mathcal{C}_k \in \mathcal{C}} \widehat{p}(\mathcal{C}_k) \prod_{j=1}^{m} \widehat{p}(x_j|\mathcal{C}_k) \tag{8}$$

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

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



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	???

Naïve Bayes Classifier: How to estimate the Probabilities?

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

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

$$\widehat{p}(x_j = v|\mathcal{C}_k) = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_j^{(i)} = v, y^{(i)} = \mathcal{C}_k\}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = \mathcal{C}_k\}} \tag{10}$$

- $\mathbb{1}\{bool\}$ is the **indicator function**

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

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

Do you have any questions?