SGN-13000/SGN-13006 Introduction to Pattern Recognition and Machine Learning (5 cr)

Bayesian Learning

Joni-Kristian Kämäräinen

September 2018

Department of Signal Processing
Tampere University of Technology

Material

- Lecturer's slides and blackboard notes
- T.M. Mitchell. *Machine Learning*. McGraw-Hill, 1997: Chapter 6
- C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006: Chapter 2
- Computer examples

Contents

Bayes theorem

Thomas Bayes: "Inverse Probability"

Figure 1: `www.york.ac.uk` .

Bayes' Theorem: A Posteriori Probability

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{1}$$

- $P(h)$ = prior probability of hypothesis h
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

Bayesian Best Hypothesis

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Generally want the most probable hypothesis given the training data

Maximum a posteriori hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i) \tag{2}$$

h_{MAP} Example

Example (Does patient have cancer or not?)

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% *of the cases in which the disease is actually present, and a correct negative result in only* 97% *of the cases in which the disease is not present. Furthermore,* .008 *of the entire population have this cancer.*

$$P(cancer) = \qquad P(\neg cancer) =$$
$$P(+|cancer) = \qquad P(-|cancer) =$$
$$P(+|\neg cancer) = \qquad P(-|\neg cancer) =$$

h_{MAP}?

Basic Formulas for Probabilities

- *Product Rule*: probability $P(A \wedge B)$ of a conjunction of two events A and B:

$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$$

- *Sum Rule*: probability of a disjunction of two events A and B:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- *Theorem of total probability*: if events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then

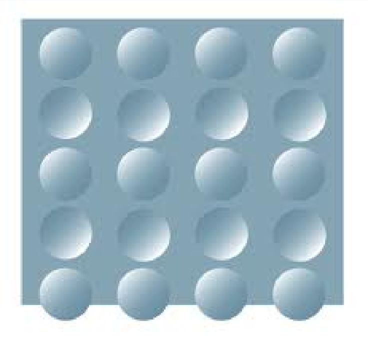
$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Prior in human cognition

Solving computer vision problems

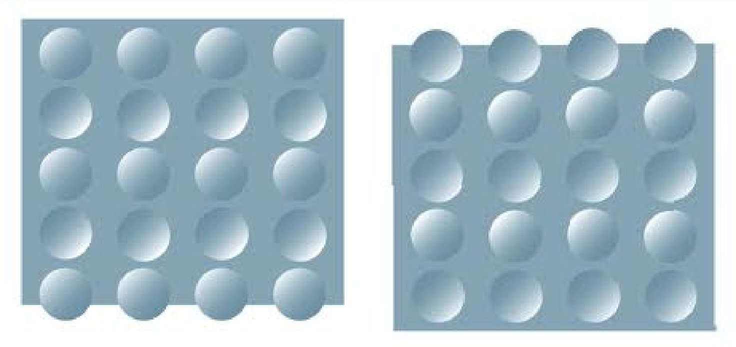
Prior in human cognition

Solving computer vision problems



Prior in human cognition

Solving computer vision problems

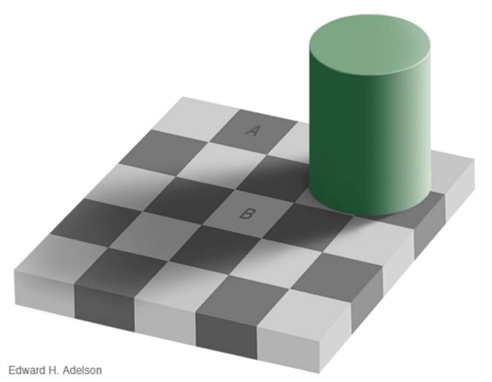


Prior in human cognition (cont.)

Prior is our experience of the physical world

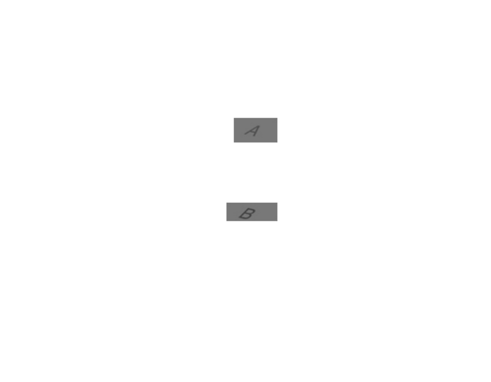
Prior in human cognition (cont.)

Prior is our experience of the physical world



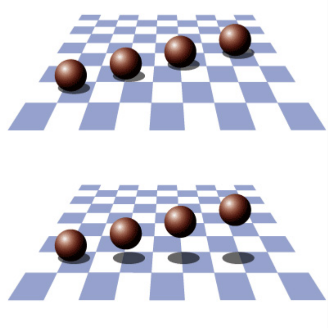
Prior in human cognition (cont.)

Prior is our experience of the physical world



Prior in human cognition (cont.)

Prior is our experience of the physical world



Be careful with probabilities

Sometimes it is better to trust algebra or experiments than intuition

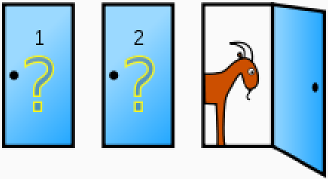


Figure 2: The Monty Hall Problem

Bayes Classifier

Most Probable Classification of New Instances

So far we've sought the most probable *hypothesis* given the data D (i.e., h_{MAP})

Given new instance x, what is its most probable *classification*?

- $h_{MAP}(x)$ is not the most probable classification!

Consider:

- Three possible hypotheses:
  $P(h_1|D) = .4, \; P(h_2|D) = .3, \; P(h_3|D) = .3$
- Given new instance x, $h_1(x) = +, \; h_2(x) = -, \; h_3(x) = -$
- What's most probable classification of x?

Bayes Classifier

Bayes optimal classifier

Bayes Optimal Classifier

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Example:

$$P(h_1|D) = .4, \quad P(-|h_1) = 0, \quad P(+|h_1) = 1$$
$$P(h_2|D) = .3, \quad P(-|h_2) = 1, \quad P(+|h_2) = 0$$
$$P(h_3|D) = .3, \quad P(-|h_3) = 1, \quad P(+|h_3) = 0$$

therefore

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D) \quad = \quad .4$$
$$\sum_{h_i \in H} P(-|h_i)P(h_i|D) \quad = \quad .6$$

and

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) \quad = \quad -$$

Bayes Classifier

Naïve Bayes classifier

Naïve Bayes Classifier

A powerful yet simple method

When to use

1. When there is no enough data points to estimate the full probabilities

Successful applications:

1. Diagnosis
2. Classifying text documents

Naïve Bayes Classifier

Assume target function $f : X \rightarrow V$, where each instance x described by attributes $\langle a_1, a_2 \dots a_n \rangle$.

Most probable value of $f(x)$ is:

$$\begin{aligned} v_{MAP} &= argmax_{v_j \in V} P(v_j | a_1, a_2 \dots a_n) \\ v_{MAP} &= argmax_{v_j \in V} \frac{P(a_1, a_2 \dots a_n | v_j) P(v_j)}{P(a_1, a_2 \dots a_n)} \\ &= argmax_{v_j \in V} P(a_1, a_2 \dots a_n | v_j) P(v_j) \end{aligned}$$

Naive Bayes assumption:

$$P(a_1, a_2 \dots a_n | v_j) = \prod_i P(a_i | v_j)$$

which gives

Naive Bayes classifier: $v_{NB} = argmax_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$

Naive Bayes Algorithm

Naive_Bayes_Learn(*examples*)

1: **for** each target value v_j **do**
2: $\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$
3: **for** each attribute value a_i of each attribute a **do**
4: $\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$
5: **end for**
6: **end for**

Classify_New_Instance(x)

$$v_{NB} = argmax_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

Naive Bayes Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Naive Bayes: Subtleties

1. Conditional independence assumption is often violated

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

   - ...but it works surprisingly well anyway. Note don't need estimated posteriors $\hat{P}(v_j | x)$ to be correct; need only that

$$argmax_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i | v_j) = argmax_{v_j \in V} P(v_j) P(a_1 \ldots, a_n | v_j)$$

2. Naive Bayes posteriors often unrealistically close to 1 or 0

Naive Bayes: Subtleties

3 what if none of the training instances with target value v_j have attribute value a_i? Then

$$\hat{P}(a_i|v_j) = 0, \text{ and...}$$
$$\hat{P}(v_j) \prod_i \hat{P}(a_i|v_j) = 0$$

Typical solution is Bayesian estimate for $\hat{P}(a_i|v_j)$

$$\hat{P}(a_i|v_j) \leftarrow \frac{n_c + mp}{n + m}$$

where

- n is number of training examples for which $v = v_j$,
- n_c number of examples for which $v = v_j$ and $a = a_i$
- p is prior estimate for $\hat{P}(a_i|v_j)$
- m is weight given to prior (i.e. number of "virtual" examples)