Exercises for
Pattern Recognition
Peter Fischer, Shiyang Hu
Assignment 1, 21/24.10.2014

SOLUTION

General Information:

Exercises (1 SWS):	Tue 12:15 – 13:45 (0.154-115) and Fri 08:15 – 09:45 (0.151-115)
Certificate:	Oral exam at the end of the semester
Contact:	peter.fischer@fau.de, shiyang.hu@fau.de

k-NN and Bayesian classification

Exercise 1 Create a k-Nearest Neighbor classifier for the Classification Toolbox, where k is a user-defined variable. Compare its output to the output of the already existing Nearest Neighbor classifier.

Hint: The k-Nearest Neighbor (k-NN) classifier calculates the distance of a new feature to the features of the training set. The new feature is assigned to the class to which the majority of the k nearest neighbors belongs to.

Exercise 2 A computer science student is annoyed that the two thirds of the e-mails he/she receives is spam. Therefore, he/she decides to write a classifier that should decide whether an incoming e-mail is spam (class y=1) or ham (class y=0). For classification the Bayes classifier is used. The student notices that in spam and ham mails, certain words occur with different probability. Therefore, the student bases the classification on the words $\boldsymbol{x} = \{Viagra, bet, student, sports, cinema\}$. By inspecting all his/her previous mails, the student estimates that in the ham mails, the probabilities are 0% for Viagra, 10% for bet, 40% for student, 30% for sports, and 10% for cinema. In the spam mails, the words Viagra occurs in 50%, bet in 30%, student in 5%, and sports and cinema in 2% of the mails.

The student does not count how often each word occurs, but only whether it is present in the mail. For simplicity, he/she assumes that the words occur independently.

(a) Has the student considered all the postulates of pattern recognition?

(a) Representative sample: the probabilities were estimated using all the previous e-mails

(b) Features: the words in the e-mails characterize the classes

(c) Compact domain: the student chose words with differing probabilities

(d) Simpler constituents: the student split the text in words and did not try to use the whole text as a feature

(e) Structure: the student did NOT consider the ordering of the words, sentence structure, ...

(f) Similarity: if the words are similar, the patterns are also similar

(b) Write down the priors for an e-mail being spam or ham.
Without looking at the text, it is known that two thirds of the e-mails are spam:
$p(y = 1) = 0.66$
$p(y = 0) = 0.34$

(c) Write down the class-conditional probabilities $p(\boldsymbol{x}|y = 0)$ and $p(\boldsymbol{x}|y = 1)$ for an arbitrary feature vector $\boldsymbol{x}$. Hint: Bernoulli distributions for each feature $p(x_i|y = 0) = p(x_i = 1|y = 0)^{x_i} \cdot (1 - p(x_i = 1|y = 0))^{1-x_i}$
Class conditional probabilities : distribution of feature vectors within a class $p(\boldsymbol{x}|y = 0)$.
Table probability distribution of each class, here class y=0 (ham)

Feature	Viagra	bet	student	sports	cinema
0 (not present)	1.0	0.9	0.6	0.7	0.9
1 (present)	0.0	0.1	0.4	0.3	0.1

mathematical formulation: Bernoulli distributions for each word

$$p(x_i|y = 0) = p(x_i = 1|y = 0)^{x_i} \cdot (1 - p(x_i = 1|y = 0))^{1-x_i}$$

Statistical independence: probabilities for different words are multiplied together

$$p(\boldsymbol{x}|y = 0) = \prod_{i=1}^{5} p(x_i = 1|y = 0)^{x_i} \cdot (1 - p(x_i = 1|y = 0))^{1-x_i}$$

$$p(\boldsymbol{x}|y = 1) = \prod_{i=1}^{5} p(x_i = 1|y = 1)^{x_i} \cdot (1 - p(x_i = 1|y = 1))^{1-x_i}$$

(d) Write down the Bayesian decision rule for the spam classification problem. Classify the following e-mail using the decision rule:
Hi, As we talked about yesterday, I want to make a bet with you about the upcoming soccer match. I clearly know more about sports than you. I bet 5$ against Nürnberg.
Bayesian decision rule: choose class with highest posterior for x. decision for class 0 if: $p(y = 0|x) > p(y = 1|x)$

$$\begin{aligned} y^* &= \operatorname*{argmax}_y p(y|\boldsymbol{x}) \\ &= \operatorname*{argmax}_y \frac{p(y) \cdot p(\boldsymbol{x}|y)}{p(\boldsymbol{x})} \\ &= \operatorname*{argmax}_y p(y) \cdot p(\boldsymbol{x}|y) \end{aligned}$$

E-mail classification:
no occurrence of viagra, student, and cinema.
one or multiple occurrences of sports and bet

feature vector $\boldsymbol{x} = \{0, 1, 0, 1, 0\}$

$$
\begin{aligned}
p(y=0) \cdot p(\boldsymbol{x}|y=0) &= p(y=0) \cdot \prod_{i=1}^{5} p(x_i = 1|y=0)^{x_i} \cdot (1 - p(x_i = 1|y=0))^{1-x_i} \\
&= 0.34 \cdot 1.0 \cdot (1 - 0.9) \cdot 0.6 \cdot (1 - 0.7) \cdot 0.9 \\
&= 0.0055 \\
p(y=1) \cdot p(\boldsymbol{x}|y=1) &= p(y=1) \cdot \prod_{i=1}^{5} p(x_i = 1|y=1)^{x_i} \cdot (1 - p(x_i = 1|y=1))^{1-x_i} \\
&= 0.66 \cdot (1 - 0.5) \cdot 0.3 \cdot (1 - 0.05) \cdot 0.02 \cdot (1 - 0.2) \\
&= 0.0018
\end{aligned}
$$

Decision: the e-mail is not spam
Hint: the probabilities do not sum to 1 because they are not normalized. The normalization is dropped in the argmax.
Hint: if you have much more features, both probabilities will become very small, which leads to numeric problems. This is one reason why the logarithm is used.

(e) Derive the decision boundary $F(\boldsymbol{x}) = \log\left(p(y=0|\boldsymbol{x})\right) - \log\left(p(y=1|\boldsymbol{x})\right) = 0$ to classify the words $\boldsymbol{x}$ and show that it is a linear function.
Requirement: Class-conditional density

$$
p(\boldsymbol{x}|y=0) = \prod_{i=1}^{5} p(x_i = 1|y=0)^{x_i} \cdot (1 - p(x_i = 1|y=0))^{1-x_i}
$$

Apply logarithm to posterior probability:

$$
\begin{aligned}
\log\left(p(y|\boldsymbol{x})\right) &= \log\left(p(y)\right) + \log\left(p(\boldsymbol{x}|y)\right) \\
\log\left(p(y|\boldsymbol{x})\right) &= \log\left(p(y)\right) + \log \prod_{i=1}^{5} p(x_i = 1|y)^{x_i} \cdot (1 - p(x_i = 1|y))^{1-x_i} \\
\log\left(p(y|\boldsymbol{x})\right) &= \log\left(p(y)\right) + \sum_{i=1}^{5} \left(x_i \log p(x_i = 1|y) + (1 - x_i) \log\left(1 - p(x_i = 1|y)\right)\right) \\
\log\left(p(y|\boldsymbol{x})\right) &= \log\left(p(y)\right) + \sum_{i=1}^{5} (x_i \left[\log p(x_i = 1|y) - \log\left(1 - p(x_i = 1|y)\right)\right] \\
&\quad + \log\left(1 - p(x_i = 1|y)\right)) \\
\log\left(p(y|\boldsymbol{x})\right) &= \log\left(p(y)\right) + \sum_{i=1}^{5} \left(x_i \log \frac{p(x_i = 1|y)}{1 - p(x_i = 1|y)} + \log\left(1 - p(x_i = 1|y)\right)\right)
\end{aligned}
$$

Boundary at equal probabilities:

$$
\begin{aligned}
F(\boldsymbol{x}) &= \log\left(p(y=0|\boldsymbol{x})\right) - \log\left(p(y=1|\boldsymbol{x})\right) = 0 \\
F(\boldsymbol{x}) &= \log\left(p(y=0)\right) + \sum_{i=1}^{5} x_i \log \frac{p(x_i=1|y=0)}{1-p(x_i=1|y=0)} + \log\left(1-p(x_i=1|y=0)\right) \\
&\quad - \log\left(p(y=1)\right) - \sum_{i=1}^{5} \left(x_i \log \frac{p(x_i=1|y=1)}{1-p(x_i=1|y=1)} + \log\left(1-p(x_i=1|y=1)\right)\right) \\
F(\boldsymbol{x}) &= \log\left(p(y=0)\right) - \log\left(p(y=1)\right) \\
&\quad + \sum_{i=1}^{5} x_i \left(\log \frac{p(x_i=1|y=1)}{1-p(x_i=1|y=1)} - \log \frac{p(x_i=1|y=0)}{1-p(x_i=1|y=0)}\right) \\
&\quad + \sum_{i=1}^{5} \left(\log\left(1-p(x_i=1|y=1)\right) - \log\left(1-p(x_i=1|y=0)\right)\right) \\
F(\boldsymbol{x}) &= A + \sum_{i=1}^{5} x_i \cdot B_i + \sum_{i=1}^{5} C_i
\end{aligned}
$$

This is a linear function in the x_i.