Exam Statistical Pattern Recognition
Friday, December 16, 2016
13.15-15.00 hours

General Instructions

1. Write your name and student number on every sheet.
2. You are allowed to use a (graphical) calculator.
3. You are allowed to consult 1 A4 sheet of paper with notes on both sides.
4. Always show how you arrived at the result of your calculations.
   Otherwise you cannot get partial credit for incorrect final answers.
5. There are five questions for which you can earn 50 points.

Question 1: Mixed Questions (11 points)

(a) (5 pnts) Suppose we want to predict the selling price (in euros) of houses. In addition to the selling price, we also have data on the lot size of the house (in square meters), and a dummy variable that indicates whether or not the house has a desirable location. We want to fit a model with the following property: houses that have a desirable location possibly have a different *price per extra square meter* than houses that do not have on a desirable location.
Which predictor variables should we include in the regression model?

(b) (3 pnts) Assume that the length of adult Dutch men and women is normally distributed with means of 182 cm and 168 cm respectively. Furthermore it is given that there is an equal proportion (50%) of man and women in the population. Somebody selects at random a person from the population and tells me the length of this person is 175 cm. If I want to minimize the probability of a wrong classification (in hypothetical repetitions of this experiment), I should predict (choose one of the options below):

(A) The person is male.

(B) The person is female.

(C) It doesn't matter, the probabilities are equal.

(D) Not enough information was provided to decide on this matter.

(c) (3 pnts) In neural networks, the back-propagation algorithm computes (choose one of the options below):

(A) The error of the gradient function.

(B) The gradient of the error function.

(C) The activations of the hidden units.

(D) The units of the hidden activations.

Question 2: Linear Regression (12 points)

Can we predict how many errors programs contain? Thomas Zimmermann and colleagues (Predicting Defects for Eclipse, Third International Workshop on Predictor Models in Software Engineering, IEEE Computer Society 2007) have tried to answer this question on a code base of the Eclipse programming environment (one of the biggest open-source projects). We analyse data of Eclipse 3.0 packages; there are 661 in total. We try to predict how many errors have been reported within 6 months after the release of a package. Call this variable `NE` (for Number of Errors). We model this problem with linear regression, using a single predictor variable, which is the total lines of code in the package divided by 100. This predictor is denoted by `TLC`. The model is estimated with the method of least squares. This produces the following results (see the extract of the R output below):

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.430274   0.150276   2.863  0.00433
TLC            0.095687   0.003082  31.051  < 2e-16

Residual standard error: 3.532 on 659 degrees of freedom
Multiple R-squared:  0.594,     Adjusted R-squared:  0.5934
F-statistic: 964.1 on 1 and 659 DF,  p-value: < 2.2e-16
```

(a) (3 pnts) According to the model, what is the expected number of errors for a package with 1000 lines of code?

(b) (3 pnts) Is the coefficient of `TLC` significant at significance level $\alpha = 0.05$?

(c) (3 pnts) What percentage of the variation in `NE` is explained by the model?

(d) (3 pnts) A potential disadvantage of the linear regression model for this application is that it might predict a negative number of errors. Could that happen with the fitted model? Explain.

Question 3: Logistic Regression (12 points)

Suppose we collect data for a group of students in a pattern recognition class with variables X_1 = hours studied, X_2 = undergraduate GPA, and $Y = 1$ if the student receives an A, and $Y = 0$ otherwise. We fit a logistic regression model and produce estimated coefficients, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, and $\hat{\beta}_2 = 1$.

(a) (4 pnts) Estimate the probability that a student who studies for 40 hours and has an undergraduate GPA of 3.5 gets an A in the class.

(b) (4 pnts) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

(c) (4 pnts) What is the marginal effect of *hours studied* (X_1) on the probability of getting an A for a student who studies for 30 hours, and has a GPA of 4?

Question 4: Support Vector Machines (9 points)

We receive the following output from the optimization software for fitting a support vector machine with linear kernel and perfect separation of the training data:

i	$x_{i,1}$	$x_{i,2}$	y_i	α_i
1	3	5	-1	0
2	4	2	-1	0
3	6	6	-1	$-\frac{2}{5}$
4	6	10	$+1$	0
5	7	8	$+1$	$\frac{2}{5}$
6	9	9	$+1$	0

Here $x_{i,1}$ denotes the value of x_1 for the i-th observation, y_i denotes the class label of the i-th observation, etc.

You are given the following formulas:

$$\beta_0 = y_s - \sum_{i=1}^{n} \alpha_i x_s^\top x_i \qquad \text{(for any support vector } x_s\text{)}$$

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i x^\top x_i$$

Answer the following questions:

(a) (3 pnts) Compute the value of the SVM bias term.

(b) (3 pnts) Give the equation of the maximum margin linear decision boundary.

(c) (3 pnts) Which class does the SVM predict for the data point $x_1 = 8, x_2 = 6$? Show your calculation.

Question 5: Maximum Likelihood Estimation (6 points)

In the lecture we have discussed the principle of maximum likelihood using a coin tossing example. In general, we want to maximize the likelihood function:

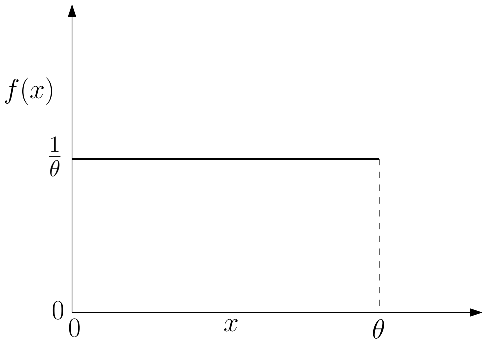
$$\ell(\theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

with respect to θ, where f denotes the density function of x, which has a parameter θ that we want to estimate.

Now let x denote a random variable with probability density function:

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

In words, x has a uniform distribution on the interval $[0, \theta]$, with $\theta \in \mathbb{R}^+$. In a picture:



The upper bound θ is unknown, and we would like to estimate it from a sample of size $n = 5$. The sample is

$$x_1 = 2.63, x_2 = 1.45, x_3 = 5.23, x_4 = 4.98, x_5 = 1.84.$$

Give the maximum likelihood estimate of θ for this sample, and explain why this is the maximum likelihood estimate. You don't have to give a formal proof; a clear informal argument suffices.