Exam Statistical Pattern Recognition
Friday, December 15, 2017
13.15-14.45 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 four questions for which you can earn 50 points.

Question 1: Mixed Questions (16 points)

(a) (4 pts) Suppose that a cholesterol lowering drug is tested through a clinical trial. We
expect there to be a linear relationship between drug dose and cholesterol reduction,
and we expect that women respond differently than men.

(1) Specify a regression equation that fits this description, where cholesterol reduc-
tion is the dependent variable, and drug dose and gender (and any variable
that may be derived from these two) are potential predictor variables. **Note:**
different specifications are defendable, so it is important that you give a solid
motivation of your choices.

(2) State whether the specification you have given under (1) satisfies the hierarchy
principle. Explain why or why not.

(b) (4 pts) For each of the following four situations, indicate whether we would gener-
ally expect the predictive performance (outside of the training sample) of a flexible
statistical learning method to be better or worse than the predictive performance
of an inflexible method. Justify your answer in each case.

(1) The sample size \( n \) is extremely large, and the number of predictors \( p \) is small.
The number of predictors $p$ is extremely large, and the number of observations $n$ is small.

The relationship between the predictors and response is highly non-linear.

The variance of the error term, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

(c) (4 pnts) In discriminant analysis, the typical assumption is that the features follow a multivariate normal distribution within each class. Different variants of discriminant analysis are obtained by imposing different constraints on the covariance matrices of these distributions. Which constraints are imposed by:

(1) Linear discriminant analysis?
(2) Naive Bayes?

(d) (4 pnts) We analyse images of $32 \times 32$ pixels with a convolutional neural network. In the first convolutional layer we create a feature map by moving a $7 \times 7$ patch across the image, starting in the upper left corner of the image. The patch is shifted each time by one pixel to the right as we move across the image. When we reach the right border of the image, we move the patch one pixel down, and start at the left border of the image again. This process continues until we reach the lower right corner of the image.

(1) How many hidden units does the feature map thus created contain?
(2) How many weights are associated with this feature map?

**Question 2: Logistic Regression (10 points)**

We analyse a data set with $28 \times 28$ pixel images of handwritten digits. Each pixel has a grayscale value between 0 (white) and 255 (black). As computer scientists we see no need to go beyond the digits 0 and 1. We extract two features from the pixel images: the sum of the pixel values divided by 1000 (ink), and the sum of the pixel values in the left half minus the sum of the pixel values in the right half, again divided by 1000 (horbal). We analyse the data with logistic regression, where digit 1 is coded as 1, and digit 0 is coded as 0. We fit the model with R, which gives the following result:

| Coefficients: | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 13.1593  | 3.1475     | 4.181   | 2.90e-05 |
| ink           | -0.6571  | 0.1547     | -4.248  | 2.16e-05 |
| horbal        | -0.7294  | 0.3052     | -2.390  | 0.0169   |
(a) (2 pts) Give an interpretation in plain language of the negative sign of the coefficient for \( \text{ink} \), that is, what does the negative sign of this coefficient mean?

(b) (2 pts) Explain why it makes sense that we found a negative sign for the coefficient of \( \text{ink} \).

(c) (2 pts) Use the fitted model to estimate the probability that an image with \( \text{ink} = 25 \) and \( \text{horbal}=0 \) contains the digit 1.

(d) (4 pts) Use the fitted model to give a linear decision rule for the classification of digits.

**Question 3: Support Vector Machines (14 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:

\[
\begin{array}{cccc}
  i & x_{i,1} & x_{i,2} & y_i \\
  1 & 3 & 4 & +1 \\
  2 & 2 & 2 & +1 \\
  3 & 4 & 4 & +1 \\
  4 & 1 & 4 & +1 \\
  5 & 2 & 1 & -1 \\
  6 & 4 & 3 & -1 \\
  7 & 4 & 1 & -1 \\
\end{array}
\]

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. The figure below plots the same data set, where circles indicate vectors of class +1, and crosses vectors of class −1.
You are given the following formulas:

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

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

Answer the following questions:

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

(b) (4 pts) Give the equation of the maximum margin linear decision boundary.

(c) (3 pts) Which class does the SVM predict for the data point $x_1 = 3, x_2 = 2$? Show your calculation.

(d) (4 pts) Suppose that after fitting the model, we found that the value of $x_1$ for the seventh observation was actually 5 instead of 4. Would this change the equation of the maximum margin linear decision boundary? Motivate your answer.

Question 4: Optimization by Gradient Descent (10 points)

A domain expert specifies the following regression function for a problem that need not concern us here:

$$Y = wX_1 + w^2X_2 + \epsilon,$$

where $Y$ is the dependent variable, and $X_1$ and $X_2$ are the predictor variables. The error term $\epsilon$ is assumed to have mean zero and constant variance $\sigma^2$. We wish to estimate the unknown coefficient $w$ by minimizing the residual sum of squares (RSS) on a data set \[ \{(y_i, x_{i,1}, x_{i,2})\}_{i=1}^{n} \] of $n$ observations.

(a) (2 pts) Give the formula for the residual sum of squares (RSS) for this regression problem.

Note that this problem is non-linear in the coefficient $w$, and there is no closed form solution for the value of $w$ that minimizes RSS. Therefore we have to resort to numerical optimization. We use the method of gradient descent to find a (local) minimum of RSS. Consider a data point with $y = 3, x_1 = 0.25, \text{ and } x_2 = 0.5$. The initial weight value $w^{(0)} = 2$. The learning rate $\eta = 0.1$.

(b) (8 pts) Compute $w^{(1)}$ by the method of gradient descent, using the given data point.