Retake Statistical Pattern Recognition
Friday, April 21, 2017
13.30-15.30 hours

General Instructions

1. Write your name and student number on every sheet.
2. You are allowed to use a 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 (10 points)

(a) (3 pnts) Consider a random experiment where two fair dice are thrown. Let X_1 and X_2 denote the outcome of the first and second throw respectively, and let Y denote the sum total of the two outcomes, that is, $Y = X_1 + X_2$. Suppose we only observe X_1, and want to predict Y. Give the values of β_0 and β_1 in the equation of the population regression line

$$E(Y \mid X_1) = \beta_0 + \beta_1 X_1$$

(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 men and women in the population, and that the standard deviation of length is the same for both groups. Somebody selects at random a person from the population and tells you the length of this person is 175 cm. If you want to minimize the probability of a wrong classification, you 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) (4 pnts) In linear model selection and regularization, give the formula for respectively the LASSO– and the ridge penalty term. Give an important consequence of the difference between the two penalty terms.

Question 2: Logistic Regression (16 points)

Judges, probation officers and parole officers are increasingly using algorithms to assess a criminal defendants likelihood of becoming a recidivist, a term used to describe criminals who re-offend. We have data on 6,172 criminal defendants from Broward County, Florida[1]. The data contains several potential predictor variables, such as gender, age, and number of prior convictions of the defendant. Furthermore it is recorded whether or not the defendant re-offended within a two-year period after release. This is coded as 0 if the person did not re-offend, and as 1 if the person re-offended. We use three predictors: gender (1 for male, 0 for female), the number of prior convictions, and age category (< 25, $25 - 45$, > 45). We estimate the model with maximum likelihood. This yields the following result (see extract from R output below):

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.040454   0.070908 -14.673   <2e-16
gender                  0.349864   0.071532   4.891    1e-06
priors_count            0.174730   0.007886  22.158   <2e-16
age > 45               -0.697008   0.075238  -9.264   <2e-16
age < 25                0.774659   0.068149  11.367   <2e-16
---

    Null deviance: 8506.4  on 6171  degrees of freedom
Residual deviance: 7603.0  on 6167  degrees of freedom
AIC: 7613
```

(a) (5 pnts) Give the estimated probability that a 30 year old male with 2 prior convictions will re-offend. Round the coefficient estimates to 3 decimals in your calculations.

(b) (3 pnts) Give a common sense interpretation of the sign of the coefficient of `gender`.

(c) (4 pnts) Based on the fitted model, give a simple classification rule to predict whether or not a male defendant below the age of 25 will re-offend within 2 years. Assume you predict the class with highest probability given the observed predictor values.

[1] (see `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`)

We apply the fitted model to the training set, and obtain the confusion matrix below (rows: predicted class, columns: true class):

	0	1
0	2,461	1,127
1	902	1,682

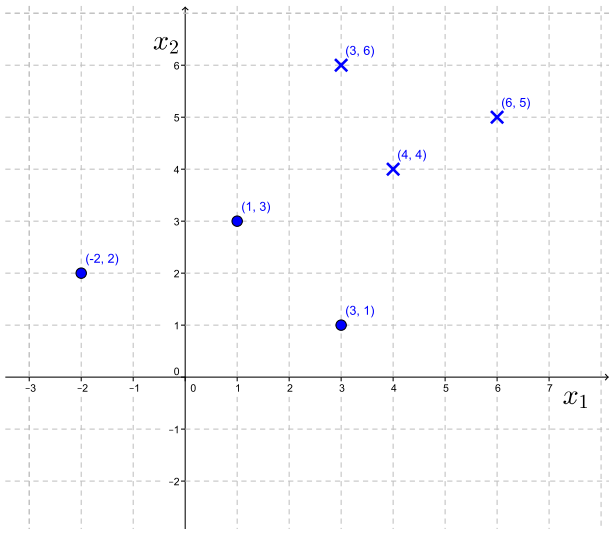
(d) (4 pnts) Give the accuracy of the classification rule, and compare it to the accuracy of the rule that simply always predicts the majority class.

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:

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

Here $x_{i,1}$ denotes the value of x_1 for the i-th observation, etc. The figure below is a plot of the same data set, where the dots represent points with class -1, and the crosses points with class $+1$.



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) There are many lines that give perfect separation of the training data. What is the defining property of the line that is preferred by the SVM algorithm?

(b) (8 pnts) Give the equation of the SVM linear decision boundary.

(c) (3 pnts) Which class does the SVM predict for the data point $x_1 = 6, x_2 = 1$? Show your calculation.

Question 4: Optimization/Gradient Descent (10 points)

In simple logistic regression, the error function is given by:

$$E(\beta_0, \beta_1) = \sum_{i=1}^{n} \left\{ y_i \ln\left(1 + e^{-\beta_0 - \beta_1 x_i}\right) + (1 - y_i) \ln\left(1 + e^{\beta_0 + \beta_1 x_i}\right) \right\}$$

where β_0 and β_1 denote the coefficients to be estimated from the data, $y_i \in \{0, 1\}$ is the value of y for the i-th observation, etc. Suppose we want to minimize this error function using the method of gradient descent.

(a) (4 pnts) Derive an expression for the gradient $\nabla E(\beta_0, \beta_1)$.
It is given that $\frac{d \ln z}{dz} = \frac{1}{z}$ and $\frac{de^z}{dz} = e^z$.

(b) (3 pnts) Let $\beta_0^{(0)} = -2$ and $\beta_1^{(0)} = 0.2$, and the step size (learning rate) $\eta = 0.1$. Use the single data point $y_i = 1, x_i = 3$ to compute the updated coefficients $\beta_0^{(1)}$ and $\beta_1^{(1)}$. **Note**: If you didn't find an answer to (a), you may assume (for partial credit) that for the given data point, the gradient evaluates to:

$$\nabla E(\beta_0, \beta_1) = \begin{bmatrix} -0.8 \\ -2.4 \end{bmatrix}$$

(c) (3 pnts) Does the gradient descent algorithm with fixed step size guarantee that

$$E\left(\beta_0^{(t+1)}, \beta_1^{(t+1)}\right) < E\left(\beta_0^{(t)}, \beta_1^{(t)}\right)?$$

Explain your answer.