# Exam Pattern Recognition
# Wednesday, April 17, 2019
# 13.30-16.30 hours

## General Instructions

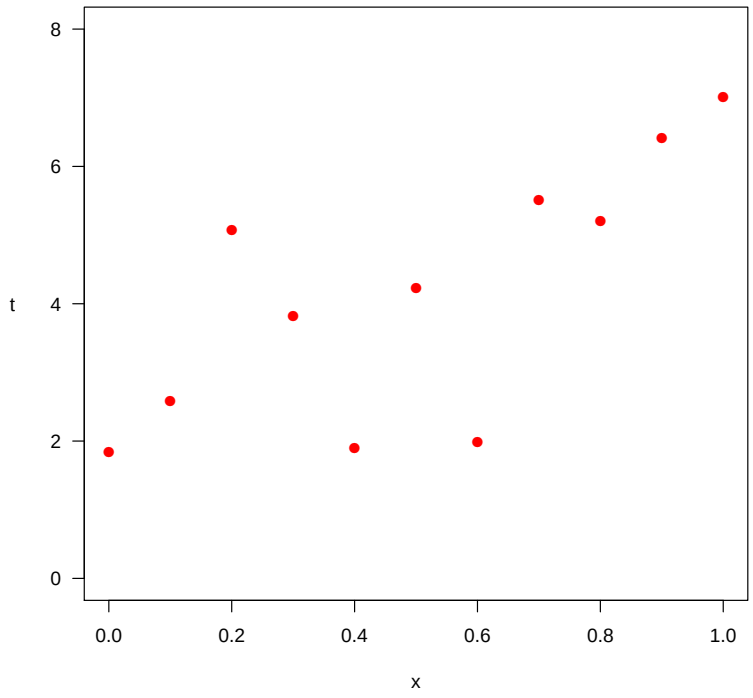
1. Write your name and student number on every sheet.
2. You are allowed to use a (graphical) calculator.
3. You are allowed to consult 1 A4 sheet of paper with notes on both sides.
4. Always show how you arrived at the result of your calculations.
   Otherwise you cannot get partial credit for incorrect final answers.
5. There are five questions for which you can earn 100 points.
6. Please write your answers to questions 1-3 and 4-5 on separate sheets. This statement is highly ambiguous, but if we tell you that questions 1,2, and 3 will be graded by Ad Feelders, and questions 4 and 5 will be graded by Zerrin Yumak, (and we would like to work in parallel) then you hopefully understand the idea.

**Question 1: General Principles (16 points)**

Suppose we generate data according to the following procedure. The predictor variable $x$ has values: $0, 0.1, 0.2, \ldots, 1$ for a total of $N = 11$ observations. The target variable $t$ is generated according to the rule

$$t_n = 2 + 4x_n + \varepsilon_n, \qquad n = 1, \ldots, 11 \tag{1}$$

where $\varepsilon_n$ is normally distributed random noise with mean 0 and standard deviation 1. An example data set generated according to this procedure is shown in the scatterplot below.




Suppose we fit two models with the following specifications to this data set (henceforth called data set A):

$$t_n = w_0 + w_1 x_n + \varepsilon_n \tag{LIN}$$
$$t_n = w_0 + w_1 x_n + w_2 x_n^2 + \varepsilon_n \tag{QUAD}$$

We use the method of least squares to estimate the unknown weights.

(a) (4 pnts) Which of the fitted models (LIN or QUAD) will have the smallest sum of squared errors on data set A? Motivate your answer.

(b) (4 pnts) Suppose we generate a fresh sample (data set B) according to the procedure stated, and use the models fitted on data set A to predict the $t$ values from the $x$ values in data set B. Which of the fitted models (LIN or QUAD) do you expect will have the smallest mean squared prediction error on B? Motivate your answer.

Alternatively, suppose the target variable $t$ is generated according to the rule

$$t_n = 2 + 4x_n - 8x_n^2 + \varepsilon_n, \qquad n = 1, \ldots, 11 \tag{2}$$

where, as before, $\varepsilon_n$ is normally distributed random noise with mean 0 and standard deviation 1. The training set is called data set C. We fit the models LIN and QUAD to data set C using the method of least squares.

(c) (4 pnts) Which of the fitted models (LIN or QUAD) will have the smallest sum of squared errors on data set C? Motivate your answer.

(d) (4 pnts) Suppose we generate a fresh sample (data set D) according to the new rule, and use the models fitted on data set C to predict the $t$ values from the $x$ values in data set D. Name the two sources of error whose sum determines whether LIN or QUAD will have the smallest mean squared prediction error on D. Discuss the trade-off between these two sources of error.

**Question 2: Logistic Regression (20 points)**

Is race a factor in denial of mortgage applications? We consider a data set compiled by researchers at the Federal Reserve Bank of Boston under the Home Mortgage Disclosure Act (HMDA). It contains data on $N = 2381$ mortgage applications filed in the Boston, Massachusetts area in 1990. The binary dependent variable is whether a mortgage application is denied (1 = denied; 0 = accepted). One important variable is the size of the required loan payments relative to the applicant's income. This variable is called `debt income ratio`: it is the ratio of the applicant's anticipated total monthly loan payments to his or her monthly income. To establish whether there is racial bias, we also include the variable `black applicant?`, which is 1 if the applicant is African-American, and 0 otherwise. We estimate a logistic regression model with the method of maximum likelihood. This yields the following result (see extract from R output below):

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.1264     0.2684 -15.372  < 2e-16
debt income ratio     5.3714     0.7284   7.374 1.65e-13
black applicant?      1.2733     0.1462   8.709  < 2e-16
```

(a) (5 pnts) Give the estimated probability that a white applicant with a `debt income ratio` of 0.33 is accepted. Round the coefficient estimates to 3 decimals in your calculations.

(b) (5 pnts) Does the sign of the estimated coefficient of `debt income ratio` make sense? Explain.

(c) (5 pnts) Based on the fitted model, give a linear classification rule to predict whether a white person's application is denied. Assume we predict the class with highest probability given the observed value of `debt income ratio`.

Do the same for a black person's application.

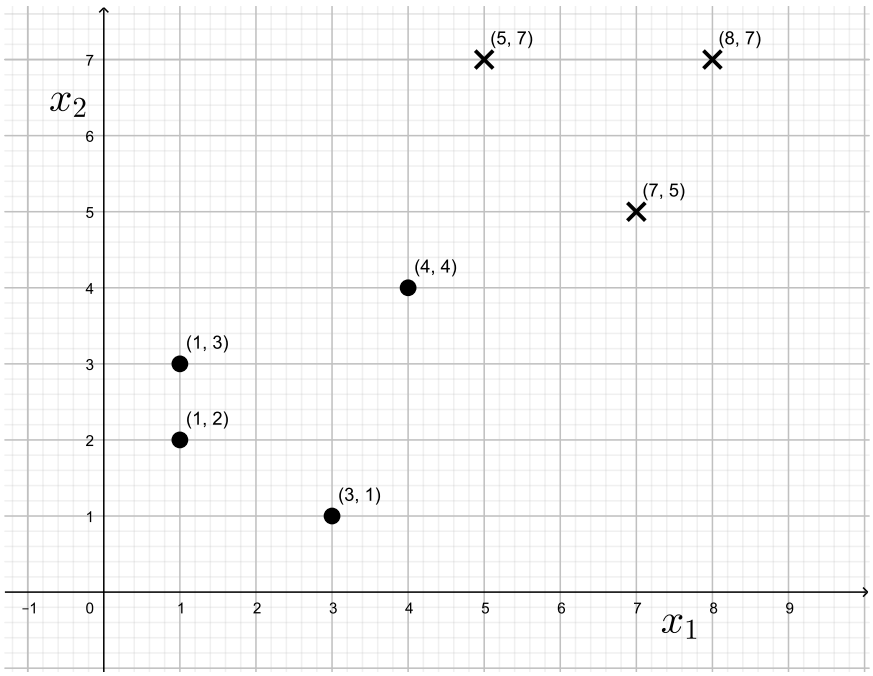
(d) (5 pnts) Does the analysis suggest that there is racial bias in the assessment of mortgage applications? Motivate your answer.

**Question 3: Support Vector Machines (24 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 ($N = 7$):

| $n$ | $x_{n,1}$ | $x_{n,2}$ | $t_n$ | $a_n$ |
|---|---|---|---|---|
| 1 | 1 | 2 | $-1$ | 0 |
| 2 | 1 | 3 | $-1$ | 0 |
| 3 | 3 | 1 | $-1$ | 0 |
| 4 | 4 | 4 | $-1$ | $\frac{1}{4}$ |
| 5 | 5 | 7 | $+1$ | $\frac{1}{8}$ |
| 6 | 7 | 5 | $+1$ | $\frac{1}{8}$ |
| 7 | 8 | 7 | $+1$ | 0 |

Here $x_{n,1}$ denotes the value of $x_1$ for the $n$-th observation, $t_n$ denotes the class label of the $n$-th observation, etc. The dataset is plotted in the figure on the next page.




You are given the following formulas:

$$b = t_s - \sum_{n=1}^{N} a_n t_n \mathbf{x}_s^\top \mathbf{x}_n \qquad \text{(for any support vector } \mathbf{x}_s\text{)}$$

$$y(\mathbf{x}) = b + \sum_{n=1}^{N} a_n t_n \mathbf{x}^\top \mathbf{x}_n$$

Answer the following questions:

(a) (4 pnts) Compute the value of the SVM bias term.

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

(c) (4 pnts) Which class does the SVM predict for the data point $\mathbf{x} = [7 \;\; 4]^\top$? Show your calculation.

(d) (4 pnts) The point $\mathbf{x}_4$ (row 4 of the data table) appears to be halfway in between the negative and positive class. Suppose we ignore (temporarily discard) this point when we determine the maximum margin decision boundary. Give the resulting decision boundary (Hint: use elementary geometrical reasoning).

(e) (4 pnts) What is the value of the slack variable $\xi_4$ for the point $\mathbf{x}_4$ based on the new decision boundary?

(f) (4 pnts) Suppose the cost parameter $C = \frac{1}{4}$. Recall that the objective function for the case with soft margin is

$$\frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{n=1}^{N} \xi_n$$

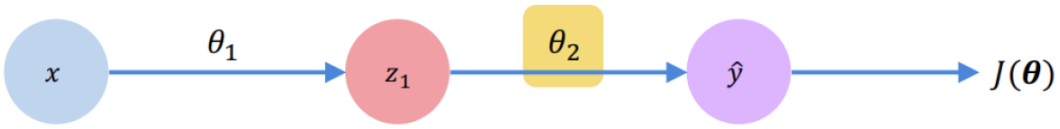
Which solution ((b) or (d)) is preferred for this value of the cost parameter?

PLEASE WRITE YOUR ANSWERS TO QUESTIONS 4 AND 5 ON A SEPARATE SHEET!

**Question 4: Gradient Descent and Backpropagation (25 points)**

(a) (8 pnts) What is the *stochastic* gradient descent algorithm and how is it different from the gradient descent algorithm? Write their pseudocode and explain the differences between them.

(b) (9 pnts) Explain how the backpropagation algorithm works using the chain rule based on the figure given below. How does a small change in weight $\theta_2$ affect the final loss $J(\theta)$ ?




(c) (8 pnts) Explain the following terminology: number of iterations, batch size, mini-batch and epoch.

**Question 5: Recurrent Neural Networks (15 points)**

(a) (4 pnts) Why are recurrent neural networks hard to train and what is the advantage of gated cells (e.g. LSTM and GRU) in comparison to plain recurrent neural networks? Explain intuitively.

(b) (6 pnts) Draw the figure of a GRU unit and write down the formula to calculate the values inside this unit.

Define each of the variables (based on the slides of Andrew Ng).

(c) (5 pnts) What is the difference between LSTM and GRU? Explain intuitively.