Exam Pattern Recognition
Friday, January 31, 2019
17.00-20.00 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 (written or printed) 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: True or False? (20 points)

For each of the statements below, indicate whether it is true or false. You don't need to explain your answer.

1. In generative classification models, one models the joint probability distribution of the features and the class label.

2. The k-means clustering algorithm is guaranteed to converge to the global minimum of its error function.

3. In linear discriminant analysis, it is assumed that the covariance matrix of the predictor variables is diagonal within each class.

4. Let $E(\mathbf{w})$ denote the error function of a learning problem with parameter vector $\mathbf{w}$. The gradient $\nabla E(\mathbf{w})$ points in the direction of steepest decrease of the error function.

5. The purpose of regularization is to prevent overfitting.

6. When trained on the same data, logistic regression and linear discriminant analysis always produce the exact same linear decision boundary.

7. Regularization tends to increase bias and reduce variance.

8. A support vector machine with non-linear kernel can always achieve perfect separation of the training data.

9. The EM algorithm is a general technique for maximum likelihood estimation for problems with latent variables.

10. In a linear regression problem with a single binary predictor variable $x \in \{0, 1\}$, the least squares estimates are $w_0 = \bar{t}_0$, and $w_1 = \bar{t}_1 - \bar{t}_0$, where $\bar{t}_0$ is the mean t value for the training examples with $x = 0$ and $\bar{t}_1$ is the mean t value for the training examples with $x = 1$.

Question 2: Logistic Regression (24 points)

We analyse a data set with 28×28 pixel images of handwritten digits. Each pixel has a grayscale value beteen 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 using maximum likelihood estimation, which gives the following result (see the next page):

Coefficient	Estimate	Std. Error	z-value	$\Pr(>\|z\|)$
(Intercept)	13.16	3.15	4.18	2.9×10^{-5}
`ink`	-0.66	0.15	-4.25	2.2×10^{-5}
`horbal`	-0.73	0.31	-2.39	0.0169

(a) (4 pnts) Give an interpretation in plain language of the negative sign of the coefficient for `ink`, that is, what does the negative sign of this coefficient mean?

(b) (4 pnts) Explain why it makes sense that we found a negative sign for the coefficient of `ink`.

(c) (4 pnts) By accident a completely white image has ended up in the test set. Use the fitted model to estimate the probability that this image contains the digit 1.

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

In the lectures, we justified the logistic regression model by listing its desirable properties. An alternative route to arrive at the logistic regression model begins as follows.
Let t^* denote a latent (unobserved) numeric variable, and assume that

$$t^* = \mathbf{w}^\top \mathbf{x} + U,$$

where U is random noise, with $\mathbb{E}[U] = 0$. Furthermore, we assume that U is symmetrically distributed around zero. As said, we can't observe the value of t^*, but we observe $t = 1$ if $t^* \geq 0$, and $t = 0$ if $t^* < 0$.

(e) (4 pnts) Show that from the assumptions above, it follows that

$$p(t = 1) = p(U \leq \mathbf{w}^\top \mathbf{x})$$

.

To continue our alternative route, assume that U follows a logistic distribution, that is, U has probability density function:

$$p(u) = \frac{e^{-u}}{(1 + e^{-u})^2}$$

(f) (4 pnts) Show that with this additional assumption, we obtain the logistic regression model. That is, show that

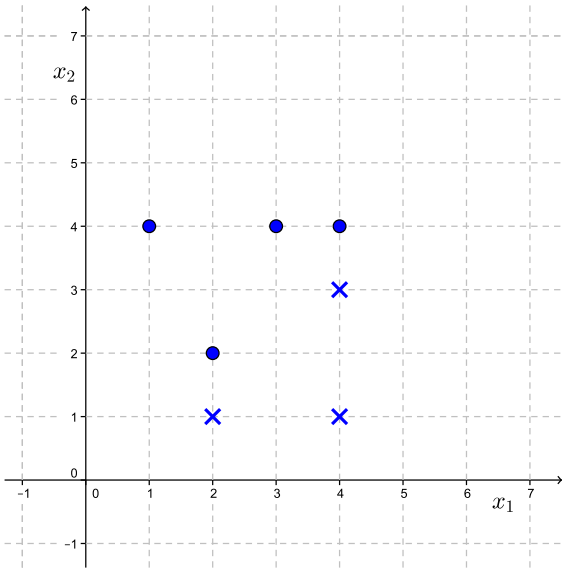
$$p(U \leq \mathbf{w}^\top \mathbf{x}) = (1 + e^{-\mathbf{w}^\top \mathbf{x}})^{-1}$$

Question 3: Support Vector Machines (20 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	$x_{n,1}$	$x_{n,2}$	t_n	a_n
1	3	4	+1	0
2	2	2	+1	2
3	4	4	+1	2
4	1	4	+1	0
5	2	1	−1	1
6	4	3	−1	3
7	4	1	−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 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:

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

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n$$

Answer the following questions:

(a) (5 pnts) Give the equation of the maximum margin linear decision boundary.

(b) (5 pnts) Give the maximum margin linear decision boundary that would be obtained if we were to remove row 2 and row 6 from the training set. Scale $\mathbf{w}$ and b appropriately. **Hint**: Use the figure and elementary geometric reasoning.

(c) (6 pnts) Recall that the error function for the case with soft margin is

$$E(b, \mathbf{w}, C) = \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C \sum_{n=1}^{N} \xi_n$$

Which solution ((a) or (b)) is preferred for $C = 1$? Take into account all training points (including row 2 and 6) when comparing the error of solution (a) and solution (b).

(d) (4 pnts) In deriving the canonical form of the SVM optimization problem for the case with perfect separation of the training data, we scaled the coefficients b and $\mathbf{w}$ such that $t_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$ for the points $\mathbf{x}_i$ closest to the decision boundary. This was translated to the set of constraints

$$t_n(\mathbf{w}^\top \mathbf{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

This set of constraints is also satisfied, however, if $t_n(\mathbf{w}^\top \mathbf{x}_n + b)$ is *strictly larger* than one for all data points. Argue that in the optimal solution this will not be the case.

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

Question 4: Convolutional Neural Networks (20 points)

(a) (6 pnts) What are the three most well-known activation functions? Write down their formula and draw their graph.

(b) (6 pnts) Explain the differences between these activation functions, what are their advantages and disadvantages?

(c) (8 pnts) You are given a CNN with two layers as shown at the top of the next page. Given an image with size $227 \times 227 \times 3$, can you define what is the output volume size and the total number of parameters after the first and second layer? Explain how you derived your calculations.

Input: $227 \times 227 \times 3$
First layer (CONV1): 96 11×11 filters are applied at stride 4
Second layer (POOL1): 3×3 filters applied at stride 2

Question 5: Recurrent Neural Networks (16 points)

(a) (6 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) (10 pnts) Draw the figure of an LSTM unit and write down the formula to calculate the values inside this unit. Define each of the variables.