Exam Contemporary Statistics

Date: Friday, January 30, 2015
Time: 9.00-12.00
Place: V 5161.0293
Progress code: WICSA-10

Rules to follow:

- This is a closed book exam. Consultation of books and notes is not permitted.
- Do not forget to fill in your name and student number.
- There are 7 exercises, and the numbers of points per exercise are indicated within boxes. 100 points can be reached and 90 points are required for the best grade (10.0); i.e. 10 points are free. The exam grade will be:

$$\text{grade} := 1 + \min\{\frac{points}{10}, 9\}$$

- We wish you success with the completion of the exam!

START OF EXAM

1. **Ridge Regression.** $\boxed{20}$
Consider a linear regression problem with p predictors and N observations:

$$y = \mathbf{X}\beta + \epsilon$$

We assume that the observations of each predictor have been standarized to mean 0 and variance 1, and that there is no intercept term β_0 in the model. The regressor matrix $\mathbf{X}$ is then an N-by-p matrix, y is the N-dimensional output vector, β is the p-dimensional vector of unknown regression coefficients, and ϵ is the N-dimensional vector of noise variables, which is here assumed to be multivariate Gaussian distributed:

$$\epsilon \sim N(0, \sigma^2 \mathbf{I}_N)$$

(a) $\boxed{5}$ For a given penalty parameter $\lambda \geq 0$ the ridge regression estimator $\hat{\beta}^{ridge}$ minimises the criterium:

$$RSS(\lambda) = (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) + \lambda \beta^T \beta$$

in β. Show that the solution is given by:

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T y$$

(b) $\boxed{4+2+4}$ Proof the following properties of the ridge regression estimator $\hat{\beta}^{ridge}$:

i. $\hat{\beta}^{ridge} = (\mathbf{I}_p + \lambda(\mathbf{X}^T\mathbf{X})^{-1})^{-1}\hat{\beta}_{LS}$, where $\hat{\beta}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$

ii. $\text{Var}(\hat{\beta}^{ridge}) = \sigma^2 (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T\mathbf{X} (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1}$

iii. $\text{Bias}(\hat{\beta}^{ridge}) = E[\hat{\beta}^{ridge}] - \beta = -\lambda(\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\beta$

to be continued below

(c) 3+2 Briefly describe how in practice an appropriate penalty parameter λ can be found: (i) Explain (see Hint) how Cross-Validation works, and (ii) explain verbally why information criteria, such as AIC and BIC, cannot be used for the determination of λ.
HINT: To explain the concept of Cross-Validation you can either give a precise verbal description or some pseudo code or a mixture thereof.

2. **Moore Penrose Pseudo-Inverse.** 20 DO^T special name

Consider a N-by-p matrix $\mathbf{X}$ with rank k, where $k \leq p \leq N$, and let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the singular value decomposition (SVD) of $\mathbf{X}$. The matrices $\mathbf{U}$ and $\mathbf{V}$ are then N-by-p and p-by-p orthogonal matrices, and $\mathbf{D}$ is a p-by-p diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$.
A matrix $\mathbf{X}^+$ is called the **Moore Penrose Inverse** of $\mathbf{X}$ if and only if the following four properties are fulfilled:
(i) $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$
(ii) $\mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+$
(iii) The matrix $\mathbf{X}^+\mathbf{X}$ is symmetric
(iv) The matrix $\mathbf{X}\mathbf{X}^+$ is symmetric

The Moore Penrose Inverse $\mathbf{X}^+$ of $\mathbf{X}$ can be computed as follows:

$$\mathbf{X}^+ = \mathbf{V}\mathbf{D}^+\mathbf{U}^T$$

where $\mathbf{D}^+$ is a p-by-p diagonal matrix with the diagonal elements $e_1, \ldots, e_p$, where $e_i = 1/d_i$ if $d_i > 0$, and $e_i = 0$ otherwise $(i = 1, \ldots, p)$.

(a) 2+2+2+2 Show that $\mathbf{X}^+ = \mathbf{V}\mathbf{D}^+\mathbf{U}^T$ fulfills the four properties (i-iv).

(b) 2 Show that the matrix $\mathbf{D}^+$ is the Moore Penrose Inverse of $\mathbf{D}$.

(c) 5 Given that the matrix $\mathbf{X}$ has full-column rank (i.e. $k = p$), show that

$$\hat{\beta} = \mathbf{X}^+ y$$

is identical to the Least-Squares (LS) estimator $\hat{\beta}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$ for the unknown regression coefficient vector β in the standard linear regression problem:

$$y = \mathbf{X}\beta + \epsilon$$

where $\mathbf{X}$ is the N-by-p regressor matrix, y is the N-dimensional output vector, and ϵ is the N-dimensional vector of noise variables, which is assumed to be multivariate Gaussian distributed: $\epsilon \sim N(0, \sigma^2 \mathbf{I}_N)$.[1]

(d) 5 Re-consider the regression problem from part (c) and briefly describe how the Bootstrap (Bootstraping procedure) could be used to approximate the covariance matrix $\text{COV}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ of the estimator $\hat{\beta} = \mathbf{X}^+ y$.
HINT: You can either give a precise verbal description or some pseudo code or a mixture thereof.

to be continued below

[1] Just note: Here it does not matter whether the columns of $\mathbf{X}$ correspond to p covariates or whether $\mathbf{X}$ is built from $p-1$ covariates with an additional column of 1's for the intercept.

3. **Cubic Spline.** 10

Consider the interval $[-2,4]$, in which we fix one point $\xi = 1$. Consider a **piecewise cubic spline** with the knot $\xi = 1$:

$$f(x) = \sum_{i=1}^{K} \beta_i h_i(x)$$

where $x \in [-2,4]$, and $h_1, \dots, h_K$ are basis functions.

(a) 5 Assume that $f(x)$ is a piecewise polynomial function with $K = 8$ basis functions. Give the basis functions $h_1, \dots, h_8$ and the three linear constraints that this cubic spline imposes on the parameters β_i $(i = 1, \dots, 8)$.

(b) 5 The same spline can also be represented with a set of truncated power basis functions, where the constraints are automatically incorperated. Represent the spline from (a) with a set of $K^\star = 5$ truncated power basis functions:

$$f(x) = \sum_{i=1}^{K^\star} \theta_i h_i^\star(x)$$

HINT: This exercise is about a 'cubic spline'; not about a 'natural cubic spline'.

4. **Piecewise-Constant Spline.** 10

Consider the interval $[-5,5]$, in which we fix two points $\xi_1 = -1$ and $\xi_2 = 2$. Consider a **piecewise constant spline** with the two knots $\xi_1 = -1$ and $\xi_2 = 2$:

$$f(x) = \sum_{i=1}^{K} \beta_i h_i(x)$$

where $x \in [-5,5]$, and $h_1, \dots, h_K$ are basis functions.
Consider the 10 data points (x_i, y_i) $(i = 1, \dots, 10)$, provided in Table 1. Use the data to fit the spline $y_i = f(x_i) + \epsilon_i$ $(i = 1, \dots, 10)$, where the noise variables are i.i.d. $N(0, \sigma^2)$ distributed, by least squares regression. That is, compute the estimator $\hat{\beta}_{LS}$ of the vector of the unknown parameters $\beta = (\beta_1, \dots, \beta_K)^T$ by plugging $\mathbf{X}$ and y into the equation:

$$\hat{\beta}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

i	1	2	3	4	5	6	7	8	9	10
x_i	-4.9	-3.2	-2.1	-1.9	-0.9	1.7	1.9	2.2	3.7	4.9
y_i	-1.0	-1.6	-0.6	-0.8	2.4	1.6	2.0	-1.5	-1.2	-0.3

Table 1: These 10 data points can be used for fitting the piecewise-constant spline.

to be continued below

5. **Linear Discriminant Analysis.** [10]
Consider a classification problem where the output Y can belong to three different classes: $Y \in \{1, 2, 3\}$. It is known that the output Y is associated with one single predictor variable X, and from training data (x_i, y_i) $(i = 1, \ldots, N)$ all unknown LDA parameters have been estimated. Thereby the following results were obtained: The estimates for the three class prior probabilities are: $\hat{\pi}_1 = 0.1$, $\hat{\pi}_2 = 0.1$ and $\hat{\pi}_3 = 0.8$, the estimates for the three class means are given by: $\hat{\mu}_1 = -2$, $\hat{\mu}_2 = 0$, and $\hat{\mu}_3 = 1$, and finally the estimate for the common variance is: $\hat{\sigma}^2 = 1$.

(a) [3] Compute the decision boundaries of this LDA model.

(b) [3] Give the resulting decision rule $\hat{y} = G(x)$.
<u>HINT</u>: Note that $G : \mathbb{R} \to \{1, 2, 3\}$ is a piece-wise function.

(c) [2] Assume that the estimated LDA model, given above, is the true underlying model. Explain verbally why the expected error rate for the classification $\hat{y}_{N+1}$ of a new observation x_{N+1} will not be zero. (One sentence might be enough.)

(d) [2] Still assuming that the LDA model is correct, give an explicit equation for the expected error rate in terms of cumulative distribution functions (CDFs).

<u>HINT</u>: A **Gaussian distribution** with parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$ has the PDF:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \cdot e^{-0.5\frac{(x-\mu)^2}{\sigma^2}}$$

for $x \in \mathbb{R}$.
The CDF of this Gaussian distribution is given by: $F_{\mu,\sigma^2}(x_0) = \int_{-\infty}^{x_0} p(x|\mu, \sigma^2)dx$

6. **AdaBoost - Population Minimizer.** [10]
Consider a binary classification problem with an output variable $Y \in \{-1, 1\}$, where $Y = -1$ means that an observation belongs to the first class, while $Y = 1$ means that an observation belongs to the second class. Let x be the realisation of a potential predictor variable X, and given $X = x$, let $\hat{y} = f(x) \in \mathbb{R}$ be a predictor for the corresponding output realisation y of Y.
Show that the predictor

$$f^\star(x) = \frac{1}{2} \log \left(\frac{P(Y = 1|X = x)}{P(Y = -1|X = x)} \right)$$

is the 'population minimiser' which minimises the conditional expectation:

$$E_{Y|X=x}[L(Y, f(x))]$$

where $L(.,.)$ is the exponential loss function with $L(a, b) = e^{-a \cdot b}$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}$.
<u>HINT</u>: For a given $x \in \mathbb{R}$, $f(x)$ is real-numbered. Hence, in your computations you might want to substitute $f(x)$ by z where $z \in \mathbb{R}$.

to be continued below

7. **EM algorithm.** [20]

In a clinical study focusing on prostate cancer for each male proband a medical diagnostic test was performed on each of 6 successive days. For $n = 196$ (diseased probands) with at least one positive test the following frequency distribution was observed:

Positive tests	0	1	2	3	4	5	6
Frequency	$Z_0 =?$	$Z_1 = 37$	$Z_2 = 22$	$Z_3 = 25$	$Z_4 = 29$	$Z_5 = 34$	$Z_6 = 49$

Table 2: **Results of the clinical study on prostate cancer.** Note that the explicit counts, provided in this table, are **not** required in this exercise.

Let the random variable Z_i describe the number of diseased probands that had i positive test results ($i = 0, \ldots, 6$), where Z_0 has not been recorded, since those probands were assumed **not** to suffer from prostate cancer.[2]

Let the random variable X_j describe the number of positive tests for proband j, and assume that the X_j's are i.i.d. and Binomial distributed with parameters $m = 6$ and π (the PDF of the Binomial distribution is given below).

(a) [5+5] Assume that the realisation of Z_0 was also known. Determine the log-likelihood $l_0(Z_0, Z_1, \ldots, Z_6; \pi)$ ('of the complete data') and derive the Maximum Likelihood (ML) estimator for π. Give all expressions in terms of the counts $Z_0, \ldots, Z_6$. That is, do <u>not</u> plug-in the concrete counts from the table.

(b) [4] Now assume that the parameter π rather than the realisation of Z_0 was known. What is then the expectation of Z_0?
To this end first determine the probability $\gamma := P(X_j = 0)$. Moreover, $Z_0 + n$ can be interpreted as the 'number of trials till $n = 196$ positive tests have been obtained'; a quantity which is negative Binomial distributed. What are the parameters of this negative Binomial distribution? And what is the conditional expectation $E[Z_0|(Z_1, \ldots, Z_6), \pi]$ of Z_0?

(c) [1] **The E-step:** Give a formula for the conditional expectation $Q(\pi, \hat{\pi}^{(j)})$, defined below, of $l_0((Z_1, \ldots, Z_6); \pi)$:

$$Q(\pi, \hat{\pi}^{(j)}) := E[l_0((Z_0, Z_1, \ldots, Z_6); \pi)|(Z_1, \ldots, Z_6), \hat{\pi}^{(j)}]$$

where $\hat{\pi}^{(j)}$ is a fixed value for π. <u>HINT</u>: Re-use your results from parts (a-b).

(d) [1] **The M-step:** Give a formula for $\hat{\pi}^{(j+1)}$ which maximises $Q(\pi, \hat{\pi}^{(j)})$ w.r.t. the free parameter π. <u>HINT</u>: Re-use your result from part (a).

to be continued below

[2] <u>Just note</u>: Even if the number of probands without any positive test result Z_0 had been recorded, it would have been the sum of those probands that actually do not have prostate cancer and those that suffer from prostate cancer but had 6 false-negative test results.

(e) $\boxed{4}$ **EM algorithm:** Give pseudo code for an EM-algorithm which iteratively infers the ML-estimator $\hat{\pi}_{ML}$ for the log-likelihood $l((Z_1,\ldots,Z_6);\pi)$ ('of the incomplete data').
HINTS: Re-use your results from the previous parts.
Proposed structure of your pseudo code:

START OF PSEUDO CODE
Initialisation: Set $\pi^{(1)} = \ldots$
Iterations For $t = 1, 2, 3$, etc.

- E-Step: Compute ...
- M-Step: Compute $\pi^{(t+1)} = \ldots$
- If ... then stop the iterations and output $\hat{\pi}_{ML} := \ldots$

END OF PSEUDO CODE

SOME GENERAL HINTS:
The density (PDF) of the **binomial distribution** with parameters $n \in \mathbb{N}$ and $\pi \in [0,1]$ is given by

$$p(x|n,\pi) = \binom{n}{x} \cdot \pi^x \cdot (1-\pi)^{n-x}$$

for $x \in \{0, 1, \ldots, n\}$.

The density (PDF) of the **negative binomial distribution** with two parameters $r \in \mathbb{N}$ and $\theta \in [0,1]$ is given by

$$p(x|\theta,r) = \binom{x-1}{x-r} \cdot (1-\theta)^{x-r} \cdot \theta^r$$

for $x \in \{r, r+1, r+2, \ldots\}$.
Also note that the expectation of the negative Binomial distribution is given by $E[X] = r/\theta$. Recall that a common interpretation is the following one: 'An experiment is successful with probability θ and it fails with the complementary probability $1-\theta$. The experiment is repeated independently. The negative Binomial distributed variable X describes how often this experiment has to be repeated until r successes have been observed.'

END OF EXAM