STATISTICAL LEARNING IN PRACTICE Part III / Lent 2020
Example Sheet 3 (of 4) By courtesy of Dr T. Wang

You have the option to submit your answers to questions 1 and 5 to be marked. If you wish your answers to be marked, please leave them in my pigeon-hole in the central core of the CMS by 11am on 11th March.

1. Consider two-class data $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}^p$, the class sizes are $N_1, N_2 \geq 1$ (so $N_1 + N_2 = n$), and y_i are coded numerically as $-n/N_1$ and n/N_2 for class 1 and 2, respectively, $i = 1, \ldots n$.

   (i) Show that LDA classifies $x \in \mathbb{R}^p$ to class 2 if

   $$x^\top \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2^\top \hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^\top \hat{\Sigma}^{-1}\hat{\mu}_1 + \log\left(\frac{N_1}{n}\right) - \log\left(\frac{N_2}{n}\right),$$

   and to class 1 otherwise.

   (ii) Consider the minimisation of the least squares criterion

   $$\sum_{i=1}^n (y_i - \alpha - x_i^\top \beta)^2.$$

   Show that the solution $\hat{\beta}$ satisfies

   $$\left\{(n-2)\hat{\Sigma} + \hat{B}\right\}\hat{\beta} = n(\hat{\mu}_2 - \hat{\mu}_1),$$

   where $\hat{B} := N_1(\hat{\mu}_1 - \hat{\mu})(\hat{\mu}_1 - \hat{\mu})^\top + N_2(\hat{\mu}_2 - \hat{\mu})(\hat{\mu}_2 - \hat{\mu})^\top$ and $\hat{\mu} = n^{-1}\sum_{i=1}^n x_i$.

   (iii) Show that $\hat{B}\hat{\beta}$ is in the direction of $(\hat{\mu}_2 - \hat{\mu}_1)$ and thus

   $$\hat{\beta} \propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1).$$

   Therefore, the least squares regression coefficient is identical to the LDA coefficient, up to perhaps a scalar multiple.

   (iv) Show that this result holds for any (distinct) numerical coding of the two class labels.

   (v) Find the solution $\hat{\alpha}$, and hence the predicted value $\hat{y}_i = \hat{\alpha} + x_i^\top \hat{\beta}$. Consider the following rule: classify the ith observation observation to class 2 if $\hat{y}_i > 0$ and class 1 otherwise. Show that this is not the same as the LDA rule unless the classes have equal numbers of observations.

   (Hastie et al., *Elements of Statistical Learning*, Exercise 4.2)

2. A researcher wants to build an email spam classifier based on a training set of $n = 500$ emails. They have hand-picked 10 words/symbols that they believe to have the highest discriminating power: `dollar`, `winner`, `password`, `edu`, `credit`, `discount`, `as`, `I`, `fun`, `trial`, and performed a logistic regression in `R`. Each row in the dataset `spamfilter` represent one email. The first column (`spam`) encodes whether an email is spam (code 1) or not (code 0) and the remaining 10 columns the count number of times a particular word/symbol appear in the email. Part of the `R` output is shown below.

```
summary(glm(spam ~ dollar + winner + password + edu + credit + discount + as + I
+ fun + trial, family = binomial)

# Coefficients:
#                    Estimate Std. Error z value Pr(>|z|)
# (Intercept)       -5.391451   1.447857  -3.724   0.0002 ***
# dollar             1.859318   1.333052   1.395   0.1631
# winner             5.680691   1.955451   2.905   0.0037 **
# password           0.923072   1.268399   0.728   0.4668
# edu               -6.890095   1.857351  -3.710   0.0002 ***
# credit             2.269523   1.007462   2.253   0.0243 *
# discount           1.198028   1.785944   0.671   0.5023
# as                -3.176676   1.832839  -1.733   0.0831 .
# I                 -1.866328   0.846315  -2.205   0.0274 *
# fun                4.347929   1.104687   2.066   0.0388 *
# trial              0.864456   1.774681   0.487   0.6262
```

(i) Write down the algebraic form of the fitted logistic regression classifier ψ^{logit} and give the estimated coefficients. How would you interpret the coefficient estimate for covariate `dollar`?

(ii) Describe algebraically the decision boundary of this classifier. In practice, one may want to only classify an email as spam if the predicted spam probability is at least a given $q \in (0, 1)$. Let ψ_q^{logit} be the corresponding classifier. Show that $\psi^{\text{logit}} = \psi_{0.5}^{\text{logit}}$. What effect does varying q have on the decision boundary?

Instead of hand-picking significant words, the researcher now wants to simply include 5000 common English words into their logistic regression classifier. The researcher fitted a ridge logistic regression model.

(iii) Explain why unregularised logistic regression is non-identifiable in this case and how the ridge regularisation resolves the non-identifiability issue.

(iv) Write down the optimisation problem solved by the ridge logistic regression algebraically. Describe how gradient descent can be performed to estimate the model coefficients.

3. Suppose we have observations $x_1, \dots, x_n \in \mathbb{R}^p$ with associated class labels $y_1, \dots, y_n \in \{-1, 1\}$. Assume that the two classes are completely separable.

(i) State the quadratic optimisation problem with linear constraints solved by the support vector machine (SVM) classifier.

(ii) Characterise the support vectors for this SVM in terms of the solution to the optimisation problem in (i).

(iii) Let $SVs = \{i : x_i \text{ is a support vector}\}$. Prove that the decision boundary of this SVM will not change if we use only $\{(x_i, y_i) : i \in SVs\}$ as our training data.

(iv) Suppose $SVs = \{1, 2\}$. Describe the decision boundary in terms of x_1, x_2. What is it if $SVs = \{1, 2, 3\}$?

4. (i) Let $\mathcal{X}$ be an arbitrary set. What is a *positive definite kernel* on $\mathcal{X}$?

   (ii) Let K_1 and K_2 are two positive definite kernels on $\mathcal{X}$ and $g : \mathcal{X} \to \mathbb{R}$ any real-valued function. Define $K_{\text{sum}}(x, x') = K_1(x, x') + K_2(x, x')$, $K_{\text{prod}}(x, x') = K_1(x, x')K_2(x, x')$ and $K_{\text{conj}}(x, x') = g(x)K_1(x, x')g(x')$. Show that K_{sum}, K_{prod} and K_{conj} are also positive definite kernels.

   (iii) Let $\mathcal{X} = \mathbb{R}^p, p \geq 1$. Using Part (ii) or otherwise, show that $K(x, x') = (c + x^\top x')^d$ and $K(x, x') = \exp\{-\gamma \|x - x'\|_2^2\}$ are both positive definite kernels for any $c > 0$, $d \in \mathbb{N}$ and $\gamma > 0$.

5. Question 3 of the 2017–2018 past paper. You may find it in the following link:

   https://www.maths.cam.ac.uk/postgrad/part-iii/files/pastpapers/2018/paper_218.pdf.

6. An advertiser wanted to understand how different web design decisions influence the effectiveness of an online advertisement. They showed the advertisement to all users visiting the website while varying the font typeface (cagegorical variable `font`, with two levels `sanserif` and `serif`), display style (variable `display`, with two levels `banner` and `popup`) and seriousness of writing (`writing`, numerical variable taking value between 1 and 10). For each user, the advertiser recorded whether the advertisement was clicked.

```
head(ad)
#   click     font display writing
# 1   yes    serif  banner       3
# 2    no sanserif  banner       8
# 3    no sanserif  banner       1
# 4    no    serif   popup       7
# 5    no    serif   popup       2
# 6    no    serif   popup       9

x <- model.matrix(~font*display, data=ad)[, -1]
y <- model.matrix(~click-1, data=ad)
layer_relu <- layer_dense(units = 2, activation = ’relu’, input_shape = dim(x)[2])
layer_softmax <- layer_dense(units = 2, activation = ’softmax’)
ad.nnet <- keras_model_sequential(list(layer_relu, layer_softmax))
compile(ad.nnet, optimizer=’sgd’, loss=’categorical_crossentropy’, metrics=’acc’)
fit(ad.nnet, x, y, batch_size=1, epochs=5)
```

   (i) Sketch a diagram of the neural network. Write down algebraically the neural network model fitted in `ad.nnet`, associating `sanserif` and `serif` to 0 and 1, `banner` and popup to 0 and 1, and no and yes to 0 and 1. Let β be the vector of all coefficients in model ad.nnet. Show that the maximum likelihood estimator for β is not unique if not restrictions are imposed on it.

(ii) Assume that all weights in the neural network are initialised to be equal to 1 and that we train the network by maximising the log-likelihood $\ell(\beta)$ via (vanilla) stochastic gradient descent with constant learning rate $\alpha = 0.1$, where α is assumed to have absorbed the sample size. What are the values of the weights after one training step in the first epoch if the data are not randomly shuffled?

(iii) If we maximise instead the regularised log-likelihood

$$\ell(\beta) - \frac{\lambda}{2}\|\beta\|_2^2$$

for some $\lambda > 0$, how would your stochastic gradient step in part (ii) change?

7. *(Exercise with R)* Download the letter dataset from the course website:

http://www.statslab.cam.ac.uk/~tw389/teaching/SLP18/data

Each row of the data contains 16 different attributes of a pixel image of one of the 26 capital letters in the English alphabet, together with the letter label itself. More information about this dataset can be found at https://archive.ics.uci.edu/ml/datasets/Letter+Recognition. Construct a classifier using one of the methods covered in this course. Then test your classifier on the test dataset letter_test from the course website. What is your test error? (Note that all design decision of your classifier must be based on the training data.)