Foundations of Data Science
Supervision 2

Andrej Ivašković (ai294)
Compiled on: 29th October 2018

YOU MAY EMAIL ME YOUR WORK OR LEAVE IT IN MY PIGEONHOLE IN THE TRINITY COLLEGE GREAT COURT MAIL ROOM.
PLEASE SUBMIT THE ASSIGNED WORK AT LEAST 24 HOURS BEFORE THE SUPERVISION!

1 Before attempting the problems

This exercise sheet covers the last bit of the course – mostly on linear regression and model fitting. There are a few questions here that reinforce the concepts from the previous parts of the course.

Linear regression is interesting in the way it combines linear algebra, optimisation and probability: it is a problem you can interpret in several different ways. Some of the techniques you use for deriving results here will be applicable in *Machine Learning and Bayesian Inference*.

2 Problems

1. A recent Ipsos MORI opinion poll[1] (fieldwork carried out 19–22 Oct 2018) investigated the popularity of UK political parties. 876 GB adults in total stated their voting preferences. Out of these, 351 people said they would vote Conservative, 324 would vote Labour, and 101 would vote for the Liberal Democrats (the remainder of people polled expresed a preference for other parties). The usual way of modelling voting preferences is by looking at every single party separately and looking at the probability a randomly drawn voter casts their vote for that party.

   (a) Calculate 95% confidence intervals for party support for the Conservatives, Labour and Lib Dems. Round your results sensibly.

   (b) Compare your result with the headline '39% Con, 37% Lab, 10% LibDem'. Is there any discrepancy, and what might be the cause of it?

   (c) How can you estimate the probability that Labour are actually ahead of the Conservatives?

[1] Source: here.

2. Attempt past exam question: 2010 Paper 8 Question 2.[2]

3. (a) What does it mean for a set of vectors to be *linearly independent*?

   (b) What are *feature vectors* and *residuals* in the context of linear models?

   (c) What does it mean for two vectors two be *orthogonal*, and for a set of vectors to be *orthonormal*?

   (d) How do we calculate the orthogonal projection of a vector onto another one? How is this result relevant for orthonormal sets of vectors?

   (e) What is the problem of *confounding variables* when fitting a model? What can we do about it?

   (f) What is *non-indentifiability* of a parameter?

4. For the stop-and-search data in section 4.1.2 of lecture notes, the proposed model was:

$$\mathbb{P}(Y_i = \text{find}) = \frac{\exp(\xi_i)}{1 + \exp(\xi_i)} \qquad \text{where} \qquad \xi_i = \alpha + \beta_{e_i} + \gamma_{g_i}$$

   where $Y_i \in \{\text{find}, \text{nothing}\}$ is the outcome of the search, g_i is the gender, and e_i is the ethnicity of suspect i. Rewrite the equation for ξ as a linear model, using one-hot coding.

5. The example given for the Iris dataset looks at the linear model:

$$\begin{bmatrix} \text{Petal.Length}_1 \\ \text{Petal.Length}_2 \\ \vdots \end{bmatrix} \approx \alpha \begin{bmatrix} 1 \\ 1 \\ \vdots \end{bmatrix} + \beta \begin{bmatrix} \text{Sepal.Length}_1 \\ \text{Sepal.Length}_2 \\ \vdots \end{bmatrix} + \gamma \begin{bmatrix} (\text{Sepal.Length}_1)^2 \\ (\text{Sepal.Length}_2)^2 \\ \vdots \end{bmatrix}$$

   Show that the set of these three basis vectors on the right hand side of the expression is not orthonormal. Can you transform it into an orthonormal system? When you do that, can you immediately get the values of α, β, γ?

   [Hint: example 5.8 in the notes illustrates the *Gram–Schmidt process* which you might want to use here.]

6. Starting from basic principles, derive closed form expressions for the maximum likelihood estimators $\widehat{a}, \widehat{b}, \widehat{\sigma}^2$ in the model $\mathbf{y} \approx a + b\mathbf{x}$ given n data points $\{(x_1, y_1), \dots (x_n, y_n)\}$ and assuming Normal$(0, \sigma^2)$ noise.

7. (a) Show that finding values a, b and c that fit $z \approx ax^b \exp(-cy^2)$, where $x_i > 0$, $z_i > 0$ for all data points $((x_i, y_i), z_i)$, can be done using linear regression. What is the assumption on the noise in that case?

   (b) Why is there no such solution for $y \approx ax^b + c$? What can we do about it instead?

   (c) What about $y \approx ax^b \exp(-cx^2)$?

[2] Yes, it is an old *Artificial Intelligence II* question. Note that you are asked about a MAP estimate – it is exactly what you expect it is, similar to the maximum likelihood approach.

8. As an alternative to the climate model given in section 5.2.2 of the lecture notes, we might suspect that temperatures are increasing linearly up to 1980, and that they are increasing linearly at a different rate from 1980 onwards. Devise a linear model to express this.

9. This question is about inference for the linear regression model:

$$\text{temp} = \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t + \text{Normal}(0, \sigma^2)$$

   (a) Give pseudocode to find the maximum likelihood estimators $\widehat{\alpha}$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\gamma}$, and $\widehat{\sigma}$.

   (b) What is meant by *parametric resampling*? Explain how to use parametric resampling to synthesize a new version of the climate dataset.

   (c) Consider the confidence interval $\gamma \in (\widehat{\gamma} \pm 0.1)$. Explain how to use bootstrap resampling to find the error probability of this confidence interval.

   (d) Give a brief outline of how to find a 95% Bayesian confidence interval for γ.

10. As hinted in the course, we can define an inner product for many kinds of vector spaces. In the vector space of real functions, we can define the inner product of two functions f and g as:

$$f \cdot g \overset{\text{def}}{=} \int_{-\infty}^{+\infty} f(x)g(x)W(x)dx$$

where W is a 'weighting' function.[3] Consider the *Legendre* definition of W:

$$W(x) \overset{\text{def}}{=} \begin{cases} 1, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Find an orthonormal basis for the vector space of third degree single-variable polynomials (with respect to this definition of inner product).

Some of the exercises have been taken from the official exercise sheet for the course. Credit for those is due to Dr Damon Wischik.

[3] You get different function families and orthogonal sets of functions for different choices of W. The Fourier basis seen in Fourier series is one such set, but there are also Jacobi polynomials, Chebyshev polynomials and others.