Example sheet 3b

Feature spaces

Foundations of Data Science—DJW—2017/2018

- You should read all the questions and understand what they are asking. You are **not** meant to answer them all. Model solutions will be provided in Easter term, and you should prepare for exams by working through the model solutions.
- You are expected to spend around 2 solid hours answering questions. Attempt questions in the order given. You should answer this example sheet with pen and paper. The answers are illustrated in notebook ex3b.ipynb.

For supervisors: This is half an example sheet. It may be supervised after the final lecture, 24 November. Model answers can be found on the course webpage.

Question 1 (one-hot coding, confounding, identifiability). In Section 3.1 of lecture notes we proposed a model for stop-and-search outcomes,

$$\mathbb{P}(Y_i = \mathsf{find}) = \beta_{e_i}$$

where e_i is the ethnicity of suspect i and $Y_i \in \{\mathsf{find}, \mathsf{nothing}\}$ is the outcome. This can be written as a linear model, using what is known as *one-hot coding*:

$$\mathbb{P}(Y_i = \mathsf{find}) = \beta_{\mathsf{Asian}} 1_{e_i=\mathsf{Asian}} + \beta_{\mathsf{Black}} 1_{e_i=\mathsf{Black}} + \cdots$$

We also proposed another model,

$$\mathbb{P}(Y_i = \mathsf{find}) = \frac{e^{\xi_i}}{1+e^{\xi_i}} \quad \text{where} \quad \xi_i = \alpha + \beta_{e_i} + \gamma_{g_i}.$$

Write the model for ξ as a linear model. Are your feature vectors linearly independent? Justify your answer. If they are not, rewrite the model in terms of a linearly independent set of feature vectors.

Question 2 (time series analysis, confounding). Let $(F_1, F_2, F_3, \dots) = (1, 1, 2, 3, \dots)$ be the Fibonacci numbers, $F_n = F_{n-1} + F_{n-2}$. Define the vectors f, f_1, f_2, and f_3 by

$$\begin{aligned} f &= [F_4, F_5, F_6, \dots, F_{m+3}] \\ f_1 &= [F_3, F_4, F_5, \dots, F_{m+2}] \\ f_2 &= [F_2, F_3, F_4, \dots, F_{m+1}] \\ f_3 &= [F_1, F_2, F_3, \dots, F_m] \end{aligned}$$

for some large value of m. If you were to fit the linear model

$$f \approx \alpha + \beta_1 f_1 + \beta_2 f_2$$

what parameters would you expect? What about the linear model

$$f \approx \alpha + \beta_1 f_1 + \beta_2 f_2 + \beta_3 f_3?$$

[Hint. Are the feature vectors linearly independent?]

Question 3 (maximum likelihood estimation). Given data $[y_1, \ldots, y_n]$, under the model $Y_i \sim \text{Normal}(\mu, \sigma^2)$ where μ and σ are unknown parameters, find the maximum likelihood estimator for σ.

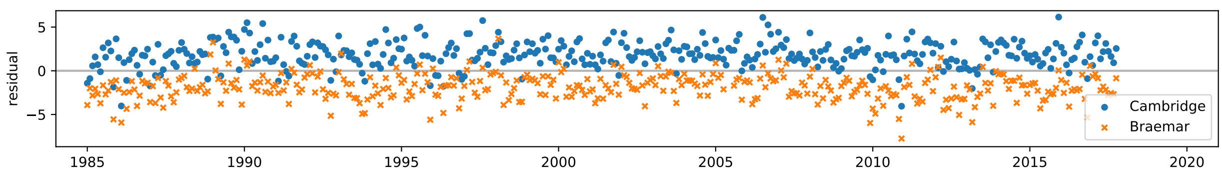
Question 4 (frequentist inference). Find a 95% confidence interval for γ, the annual rate of temperature increase for Cambridge station, from the model in Section 5.2 of lecture notes. You should give pseudocode and explain your reasoning carefully; you need not actually program anything. *[Hint. Read the note about parametric resampling in Section 5.4.]*

Question 5 (residuals, one-hot coding). It is often illuminating to plot the residual vector, to find out if we have missed any features worth including. In a probabilistic linear regression model (as in Section 5.4), the residual vector should consist of independent $\text{Normal}(0, \sigma^2)$ random variables. The residual plot should not show any systematic patterns nor signs of non-normality; if it does then we should try a different model.

Consider the two weather stations Cambridge and Braemar, plotted in Section 5.2 of lecture notes, and suppose we fit the model

$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma \mathsf{t}$$

to the dataset consisting of records from both those stations. Here is a plot of the residuals. Explain what you see. Suggest an improvement to the model.



Question 6 (contrasts). Consider the temperature data for Cambridge, from Section 5.4. Here are two models:

$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma \mathsf{t}, \tag{1}$$

and

$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma (\mathsf{t} - 2000).$$

The first model produces a fitted value $\alpha = -63.9$°C and a 95% confidence interval $[-96.5, -34.7]$°C. The second model produces a fitted value $\alpha = 10.5$°C and a 95% confidence interval $[10.4, 10.7]$°C. Why the difference? Why is the confidence interval much smaller in the second case? Which is correct?

Question 7 (linearity of trends, factors, one-hot coding). For the climate data from Section 5.2 of the notes, we proposed the model (1), in which the term $+\gamma \mathsf{t}$ asserts that temperatures are increasing at a constant rate.

Suppose we create a non-numerical feature out of t by

u = 'decade_' + str(math.floor(t/10)) + '0s'

(which gives values like 'decade_1980s', 'decade_1990s', etc.), and fit the model

$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma_{\mathsf{u}}$$

Write this as a linear model. Explain how we might use it to investigate whether temperatures are indeed increasing at a constant rate. What are the advantages and disadvantages of this model, as opposed to fitting the model (1) separately for each decade?

[Python and numpy do not have good support for enum types in the way Java does, so this code stores u *as a string. In data science, enum features are called* factors *or* categorical variables.*]*