Example sheet 3

Model crafting

Foundations of Data Science—DJW—2018/2019

Question 1. For the stop-and-search data in section 4.1.2 of lecture notes, we proposed a model

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

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

Question 2. For the climate data from section 5.2.2 of lecture notes, we proposed the model

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

in which the $+\gamma\mathsf{t}$ term asserts that temperatures are increasing at a constant rate. We might suspect though that temperatures are increasing non-linearly, as discussed in section 5.2.3. To test this, we can create a non-numerical feature out of t by

$$\mathsf{u} = \text{'}\mathsf{decade\_}\text{'} + \mathsf{str(math.floor(t/10))} + \text{'}\mathsf{0s}\text{'}$$

(which gives us 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, and give pseudocode to fit it. *[You should explain what the feature vectors are, then give a one-line command to estimate the parameters.]*

What are the advantages and disadvantages of this model, as opposed to fitting (1) separately for each decade?

Question 3. As an alternative to the climate model (1), 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.

Question 4. This question is about inference for the linear regression model

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

(a) Give pseudocode to find the maximum likelihood estimators $\hat\alpha$, $\hat\beta_1$, $\hat\beta_2$, $\hat\gamma$, and $\hat\sigma$.

(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 $\gamma \in \hat\gamma \pm 0.1$. 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.

Question 5. Here are two different models for the climate data:

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

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 6. In your answer to question 1, are your feature vectors linearly independent? Justify your answer. If not, rewrite the model in terms of a linearly independent set of feature vectors.

Question 7. 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 8. Three chess players play each other. In a tournament, A won 7 matches against B and lost 3, A won 9 matches against C and lost 1, and B won 6 matches against C and lost 4. We wish to ascribe a skill level to each player, such that the higher the skill difference the more likely it is that the higher-skilled player wins a match. Let μ_A, μ_B, and μ_C be skill levels, and consider this model: if match i is between players $p1(i)$ and $p2(i)$ then the probability that $p1(i)$ wins is $e^{\xi_i}/(1 + e^{\xi_i})$ where $\xi_i = \mu_{p1(i)} - \mu_{p2(i)}$.

(a) Find the log likelihood of (μ_A, μ_B, μ_C)

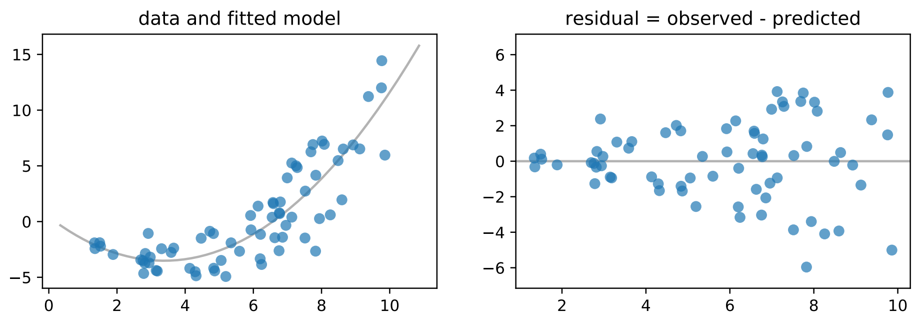
(b) Show that these parameters are not identifiable, and give an equivalent 'reduced' parameterization that is identifiable.

[Hint. This is like question 6.] In IA Algorithms we learnt the topological sort algorithm, which puts items in order given a set of pairwise comparisons. That algorithm only works with perfect non-noisy data, whereas machine learning models like this chess skill model can cope with noise.

Question 9. We are given a dataset (`https://teachingfiles.blob.core.windows.net/founds/ex3q9.csv`) with covariate x and response variable y and we fit the linear model

$$y_i \approx \alpha + \beta x_i + \gamma x_i^2.$$

After fitting the model using the least squares estimation, we plot the residuals $\varepsilon_i = y_i - (\hat{\alpha} + \hat{\beta} x_i + \hat{\gamma} x_i^2)$.



(a) Describe what you would expect to see in the residual plot, if the assumptions behind linear regression are correct.

(b) This residual plot suggests that perhaps $\varepsilon_i \sim \text{Normal}\big(0, (\sigma x_i)^2\big)$ where σ is an unknown parameter. Assuming this is the case, give pseudocode to find the maximum likelihood estimators for α, β, and γ.