Week 5 exercises

This is the third page of *assessed* questions, as described in the background notes. These questions form 70% of your mark for Week 5. The introductory questions in the notes and the Week 5 discussion group task form the remaining 30% of your mark for Week 5.

Unlike the questions in the notes, you'll not immediately see any example answers on this page. However, you can edit and resubmit your answers as many times as you like until the deadline (Friday 23 October 4pm UK time). This is a *hard deadline*: This course does not permit extensions and any work submitted after the deadline will receive a mark of zero. See the late work policy.

Queries: Please don't discuss/query the assessed questions on hypothesis until after the deadline. If you think there is a mistake in a question this week, please email Arno.

Please only answer what's asked. Markers will reward succinct to-the-point answers. You can put any other observations in the "Add any extra notes" button (but this is for your record, or to point out things that seemed strange, not to get extra credit). Some questions ask for discussion, and so are open-ended, and probably have no perfect answer. For these, stay within the stated word limits, and limit the amount of time you spend on them (they are a small part of your final mark).

Feedback: We'll return feedback on your submission via email by ~~Wednesday 28 October (sorry)~~ Friday 30 October.

Good Scholarly Practice: Please remember the University requirements for all assessed work for credit. Furthermore, you are required to take reasonable measures to protect your assessed work from unauthorised access. For example, if you put any such work on a public repository then you must set access permissions appropriately (permitting access only to yourself). You may not publish your solutions after the deadline either.

1 Basis functions and regression

1. **Radial Basis Functions (RBFs):** (40 marks)

   In this question we form a linear regression model for one-dimensional inputs: $f(x) = \mathbf{w}^\top \boldsymbol{\phi}(x; h)$, where $\boldsymbol{\phi}(x; h)$ evaluates the input at 101 basis functions. The basis functions

   $$\phi_k(x) = e^{-(x-c_k)^2/h^2}$$

   share a common user-specified bandwidth h, while the positions of the centers are set to make the basis functions overlap: $c_k = (k - 51)h/\sqrt{2}$, with $k = 1 \ldots 101$. The free parameters of the model are the bandwidth h and weights $\mathbf{w}$.

   The model is used to fit a dataset with $N = 70$ observations each with inputs $x \in [-1, +1]$. Assume each of the observations has outputs $y \in [-1, +1]$ also. The model is fitted for any particular h by transforming the inputs using that bandwidth into a feature matrix Φ, then minimizing the regularized least squares cost:

   $$C = (\mathbf{y} - \Phi\mathbf{w})^\top (\mathbf{y} - \Phi\mathbf{w}) + 0.1\mathbf{w}^\top \mathbf{w}.$$

   Hint: The question is easier to answer if you sketch the arrangement of the basis functions and mark the $[-1, +1]$ range of the inputs in your sketch. This sketch is not for credit or submission, but to help you understand the question.

   a) *[The website version of this note has a question here.]*

   b) *[The website version of this note has a question here.]*

2. **Multiple regression models:** (30 marks)

   We have a dataset of inputs and outputs $\{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, describing N preparations of cells from some lab experiments. The output of interest, $y^{(n)}$, is the fraction of cells that are alive in preparation n. The first input feature of each preparation indicates whether the cells were created in lab `A`, `B`, or `C`. That is, $x_1^{(n)} \in \{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$. The other features are real numbers describing experimental conditions such as temperature and concentrations of chemicals and nutrients.

   The lab identity is a categorical variable. A standard way to encode such variables is a 1-of-M, or 'one-hot' encoding. The feature is replaced with $M = 3$ binary features. All but one of the features are set to zero, and one of them is set to one, indicating whether the lab was `A`, `B`, or `C`. In this question, do not worry about the three binary features being linearly dependent.

   You want to predict the fraction of alive cells in future preparations from these labs using linear regression without basis functions. Compare using the lab identity as an input to your regression (as described above), with two baseline approaches: i) Ignore the lab feature, treat the data from all labs as if they came from one lab; ii) Split the dataset into three parts one for lab `A`, one for `B`, and one for `C`. Then train three separate regression models.

   a) *[The website version of this note has a question here.]*

   b) *[The website version of this note has a question here.]*

   c) *[The website version of this note has a question here.]*