EE-UY 4423: Introduction to Machine Learning
Midterm 1, Fall 2016

Answer all THREE questions. Exam is closed book. No electronic aids. But, you are permitted a limited number of cheat sheets. Part marks are given. If you do not remember a particular python command or its syntax, use pseudo-code and state what syntax you are assuming.

Best of luck!

1. A sports scientist wants to estimate the heart rate, y, immediately after some number, x of minutes of exercise. She considers a simple linear model of the form,

$$\hat{y} = ax + b, \tag{1}$$

where $\hat{y}$ is the predicted heart rate (in beats per minute or bpm), and a and b are parameters to be estimated. She gets five data points (x_i, y_i)

Exercise (minutes)	x_i	0	0	1	2	3
Heart rate (bpm)	y_i	75	65	90	110	130

(a) What are the units of a and b in the model?

(b) Show exactly how you would calculate the least-squares fit for a and b. You do not need to work out all the calculations, but make sure that enough details are given that one could work it out.

(c) Eye-balling the data, approximately what are reasonable values for a and b? Since you are not using a calculator, any reasonable value is OK.

(d) Using the values of a and b in part (c), what would the model predict for the heart rate after 30 minutes of exercise. Is this reasonable?

(e) Write a few lines of python code to plot a scatter plot of the five data points above. On the same plot, your code should also plot the line (1) for given values `a` and `b`. You can assume you have loaded the python libraries:

```
import matplotlib
import matplotlib.pyplot as plt
```

2. An engineer wishes to understand power consumption in smart phones. He collects data as shown in Table 1. He then tries to fit a model of the form,

$$\hat{y} = f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

where

- $\hat{y}$ is the predicted power in mW
- $x_1 = 0$ if the display is off and x_1 is the screen area if the display is on;
- x_2 is the data rate in Mbps;
- x_3 is the CPU usage.

The brand name and data service (cellular or WiFi) are, at first, ignored in the model.

Brand	Screen area (cm^2)	Display	Data rate (Mbps)	Data service	CPU usage	Power (mW)
Samsung G6	110	Off	0.01	WiFi	0.1	100
iPhone 7+	120	On	1	WiFi	0.8	400
iPhone 6	92	On	3	Cellular	0.6	700
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Smart phone power consumption data. Only the first three of $N = 100$ data records are shown.

(a) Let $\mathbf{A}$ be the transformed data matrix such that $\hat{\mathbf{y}} = \mathbf{A}\boldsymbol{\beta}$, where $\hat{\mathbf{y}}$ is the vector of predicted power consumption values on the training data. Let $\mathbf{y}$ be the vector of measured power consumption values in the training data. Using the values in Table 1 as training data, write the first three rows of $\mathbf{A}$ and $\mathbf{y}$.

(b) Suppose that least-squares fit on the training data yields parameter estimates

$$\hat{\boldsymbol{\beta}} = [50, 2, 25, 300],$$

with appropriate units. What is the predicted value, $\hat{y}$ at a value $\mathbf{x}$ corresponding to the case where the display is off, the CPU usage at 0.5 and data rate at 1 Mbps?

(c) Suppose that the "true" power consumption is given by

$$y = f_0(\mathbf{x}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \quad \sigma_\epsilon = 100 \text{ mW}, \tag{2}$$

where ϵ is Gaussian error with standard deviation and $f_0(\mathbf{x})$ is "true" functional relationship between the variables $\mathbf{x}$ and power consumption y. Suppose the $N = 100$ training samples are generated by the model (2), and the inverse autocorrelation of the feature vector $\phi(\mathbf{x}) = [1, x_1, x_2, x_3]^\mathsf{T}$ is given by

$$R_{\phi\phi}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 10^{-4} & 0 & 0 \\ 0 & 0 & 0.1 & 0.02 \\ 0 & 0 & 0.02 & 2 \end{bmatrix}.$$

If there is no under-modeling, what would the mean-squared error (MSE) in the prediction $\hat{y}$ from the previous part be

$$\text{MSE}(\hat{y}) = \mathbb{E}\left(\hat{y} - y\right)^2.$$

(d) Now, as a refinement to the model, suppose that the engineer wants to use a different slope for the data rate depending on whether the data is via cellular or WiFi. How would you modify the transformed features? Rewrite the first three rows for the matrix $\mathbf{A}$ in part (a) using the new model.

3. A data scientist at a retailer wishes to model the effects of advertising on sales and the duration of those effects. She considers a model of the form,

$$\hat{z}_k = a + bx\rho^k, \tag{3}$$

where $k = 0, 1, 2, \ldots$ is the index of the day, starting with $k = 0$ on the first day after the advertising purchase; $\hat{z}_k$ is the predicted value of sales (in 1000s of dollars) on the k-th day after the ad purchase; x is amount of money spent on the ad purchase (in 1000s of dollars); and $\rho = 0.9$ is a fixed constant. The parameters a and b are unknown and to be estimated. To fit the model, she collects sales records following two advertising purchases:

- Ad purchase 1: $x =$ \$10,000 were spent and sales were measured on five days ($k = 0, 1, \ldots, 4$) following the ad purchase; and
- Ad purchase 2: $x =$ \$20,000 were spent and sales were measured on ten days ($k = 0, 1, \ldots, 9$) following the ad purchase.

Thus, the data set has a total of $N = 15$ samples that can be used for training the model.

(a) Find a parameter vector $\boldsymbol{\beta}$ and transform feature vector $\phi(x, k)$ such that the predicted sales is given by $\hat{z}_k = \phi(x, k)^\mathsf{T}\boldsymbol{\beta}$.

(b) Describe the feature matrix $\mathbf{A}$ such that $\hat{\mathbf{z}} = \mathbf{A}\boldsymbol{\beta}$, where $\hat{\mathbf{z}}$ is the vector of recorded sales. You do not need to work out any exponents in your answer, just provide enough details that the components of $\mathbf{A}$ can be precisely worked out.

(c) For the remainder of the problem, suppose that the true relation between sales and ad purchases is given by

$$z_k = f_0(x, k) + \epsilon, \quad f_0(x, k) = a_0 + b_0 x \rho_0^k, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{4}$$

for some "true" parameters a_0, b_0 and ρ_0, and model standard deviation σ_ϵ. Under what condition on the "true" parameters a_0, b_0 and ρ_0 does the model (3) introduce no under-modeling with respect to the true model (4).

(d) Let $\hat{\boldsymbol{\beta}}$ be the least-squares estimate of the parameters using the $N = 15$ training samples described above with the model (3). The data scientist uses the estimated parameters $\hat{\boldsymbol{\beta}}$ to make a prediction, $\hat{z}_k$ of the sales $k = 2$ days after an ad purchase of $x =$ \$30,000. From class, we know the expected prediction (averaging over the noise) is given by

$$\bar{z}_k = \mathbb{E}(\hat{z}_k) = \phi(x, k)^\mathsf{T} R_{\phi\phi}^{-1} R_{\phi f_0},$$

where

$$R_{\phi\phi} = \frac{1}{N}\mathbf{A}^\mathsf{T}\mathbf{A}, \quad R_{\phi f_0} = \frac{1}{N}\mathbf{A}^\mathsf{T}\mathbf{f}_0,$$

and $\mathbf{f}_0$ is the vector of true function values $f_0(x, k)$ on the training data. Describe the vector $\mathbf{f}_0$. Also, using the above formula, write a few lines of python code to compute $\bar{z}_k$ at $x = 30$ and $k = 2$. You **do not** need to include the construction of the matrix $\mathbf{A}$ and $\mathbf{f}_0$ in your code. You can assume $a_0 = 50$, $b_0 = 0.7$, $\rho_0 = 0.8$ and $\rho = 0.9$.

(e) Now suppose you needed to estimate to estimate ρ in the model (3) along with a and b. Can you think of a method to estimate all three parameters, a, b and ρ? There is no single correct solution. Write your answer in pseudo-code or python.