EE-UY/CS-UY 4563: Introduction to Machine Learning
Midterm 1, Fall 2017

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 researcher wishes to evaluate a drug that claims to decrease the heart rate of cardiac patients. She decides to collect data and fit a model,

$$\hat{y} = f(x),$$

where $\hat{y}$ is the predicted decrease in heart rate in beats per minute (bpm) and x is the drug dosage in milligrams (mg). She collects data as shown in Table 1. Note that only the first three samples are shown.

Patient ID	Heart rate (bpm)		Dosage (mg)
	Before drug	After drug	
123	80	70	20
456	90	75	70
789	85	55	120
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 1: Training data for the model. Only the first three data records are shown.

(a) She first hypothesizes the model $f(x)$ that is quadratic in x (i.e. depends on x and x^2). But, she wants to force the condition when $x = 0$, $\hat{y} = 0$. Write a mathematical equation for $f(x)$ and identify the unknown parameters in that model.

(b) After more review, she comes to believe that dosages above $x \geq 100$ mg have no effect. She thus considers an alternate model where $f(x)$ is a quadratic function for x for $x < 100$ mg and $f(x)$ is constant for $x \geq 100$ mg. She still wants to force the condition that $f(0) = 0$. She also wants to force that $f(x)$ is continuous in x (that is there is not a discontinuity at $x = 100$). Write a mathematical equation for $f(x)$ and identify the unknown parameters in that model. Ensure that your model is linear in the parameters.

(c) Since the model in part (b) is linear in the parameters, we should be able to write $\hat{\mathbf{y}} = \mathbf{A}\beta$, where $\hat{\mathbf{y}}$ is a vector of predicted values, β is a vector of coefficients and $\mathbf{A}$ is some matrix depending on the data $\mathbf{x}$. Given the data in Table 1, write the first three rows of the matrix $\mathbf{A}$ and the first three values of measured response vector $\mathbf{y}$.

(d) Suppose you are given python vectors `dosage`, `hr_before` and `hr_after` representing the dosages, and heart rates before and after the drug on a set of trials. Write a python function to create a data matrix `Atr` and response vector `ytr` using the first 50 trials of the experiment. You can use the python function,

```
A = np.column_stack((col1,col2,...,coln))
```

that creates a matrix `A` with columns `col1,...,coln`.

2. Suppose we try to fit a linear model of the form,

$$\hat{y} = f(x, \boldsymbol{\beta}) = \beta_1 x + \beta_0, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ are unknown parameters. But, the true data is exponential,

$$y = f_0(x) = Ce^{-Dx} \tag{2}$$

for some values $C > 0$ and $D > 0$.

(a) For what values of $C > 0$ and $D > 0$, if any, is the true function $f_0(x)$ in (2) in the model class $f(x, \boldsymbol{\beta})$ in (1).

(b) The model is trained on data (x_i, y_i), $i = 1, \ldots, n$. Suppose that all the values x_i are close to some nominal value, x_{nom}, so that a linear approximation of the true function is valid,

$$y_i = f_0(x_i) \approx f_0(x_{\text{nom}}) + f_0'(x_{\text{nom}})(x_i - x_{\text{nom}}).$$

Using the model (1), what would we obtain for the least-squares estimate $\hat{\boldsymbol{\beta}}$? Write $\hat{\boldsymbol{\beta}}$ in terms of C, D and x_{nom}. Assume that the least-squares estimate is unique.

Hint: When the true data is linear and there is no noise, linear estimation recovers the true parameters for any set of training data when the least-squares is unique.

(c) Suppose that $C = 1$, $D = 0.5$ and $x_{\text{nom}} = 2$. Approximately draw the true function $f_0(x)$ and estimated function $f(x, \hat{\boldsymbol{\beta}})$. Your drawing does not have to be exact, but label the point x_{nom}. At what point do you think the bias would be largest, $x = 2, 3$ or 4?

(d) Suppose now that we consider higher order polynomial models to fit the data. You are given functions:

```
betahat = fit(x,y,d)
yhat = predict(x,beta)
```

where `fit` fits a polynomial model of degree `d` using training data `x` and `y`, and `predict` computes the predicted values of a polynomial with parameters `beta` and values `x`. Using these two functions, write python code that

- Given data `x` and `y`, splits the data into 50 training samples and the remaining samples being for test.
- Fits models for model orders `d=1,2,...,5`.
- Selects the model order with the lowest test RSS.

Note: You do not need to implement the one SE rule. Just pick the lowest test RSS. Also, you do not need to do K-fold validation. Just validation with one training and test split.

3. A researcher wishes to find when our eyes can detect flashes of light. She conducts a number of trials, where in each trial she shines a light of power P and at angle θ and records whether the person saw the light ($y = 1$) or did not see the light ($y = 0$). She hypothesizes that she can build a linear predictor using the variables

$$\mathbf{x} = (x_1, x_2) = (P\cos(\theta), P\sin(\theta)).$$

The first four data points of her experiments are shown in the table below.

Trial	Power P	Angle θ (deg)	Detected? (0=no or 1=yes)
1	1	0	0
2	3	45	1
3	4	85	0
4	3	-45	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$

(a) Draw a scatter plot of the points (x_{i1}, x_{i2}) using a different marker for the detected and undetected points. You do not need to calculate $\cos(\theta)$ or $\sin(\theta)$ precisely for the points. Just draw by hand where you approximately think they should be.

Hint: Consider polar coordinates.

(b) Are the first four data points linearly separable? If so, provide at least one classifier that is linear in $\mathbf{x}$ that separates the two classes. If not, find a classifier that makes the least number of errors on the first four points.

(c) Suppose that after more training data, the researcher fits a logistic model,

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

with coefficients $\boldsymbol{\beta} = (1, 2, 3)$. Based on this model, for a given power $P > 0$, what value of θ maximizes the probability of detection?

(d) Write a python function `predict` that takes vectors of power values and angles as well as the parameters $\boldsymbol{\beta}$ and returns the class probabilities $P(y_i = 1|\mathbf{x}_i)$ for each sample i.