Lab 10
Machine Learning

A. Multiple Choice (35 points, 5 points each question)

1. Machine learning (ML) means...

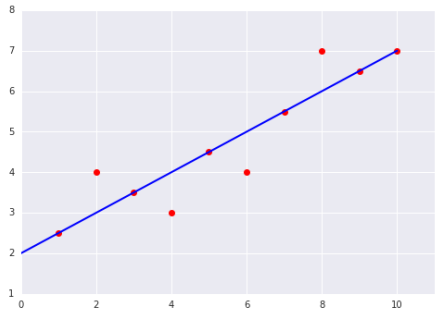
   (a) Training a statistical model based on example data
   (b) Teaching people how to use machines better
   (c) Copying how human experts do tasks
   (d) Programming a set of logical rules to perform a task

2. What is the output of a regression?

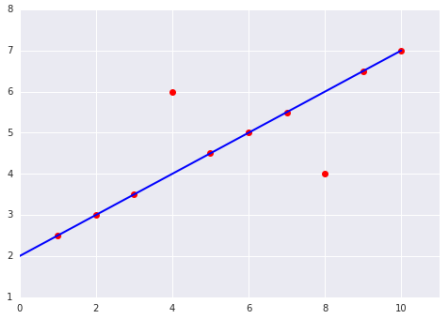
   (a) A reward or punishment
   (b) One of a set of categories
   (c) One or more numbers
   (d) A picture

3. Which of the two data sets shown in the following plots has the higher Mean Squared Error (MSE)?

   (a)



   (b)



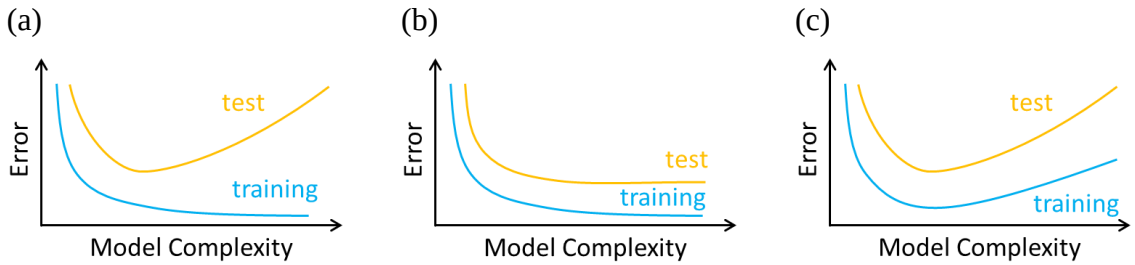
4. Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting?

   (a) Increase the amount of training data
   (b) Improve the optimization algorithm being used for error minimization
   (c) Decrease the model complexity
   (d) Reduce the noise in the training data

5. The following table illustrates the result of evaluating 4 models with different parameter choices on same dataset. Which of the following models fits this data the best?

Model	Parameter (intercept, slope)	Mean Squared Error (MSE)
(a)	(0.0, 1.4)	20.51
(b)	(3.1, 1.4)	15.23
(c)	(2.7, 1.9)	13.67
(d)	(0.0, 2.3)	18.99

6. Which of the following plots would you NOT expect to see as a plot of training and test error curves?



7. Suppose you have picked the parameter θ for a model using 10-fold cross validation. The best way to pick a final model to use and estimate its error is to...
   (a) pick any of the 10 models you built for your model; use its error estimate on the held-out data
   (b) pick any of the 10 models you built for your model; use the average CV error for the 10 models as its error estimate
   (c) average all of the 10 models you got; use the average CV error as its error estimate
   (d) average all of the 10 models you got; use the error the combined model gives on the full training set
   (e) train a new model on the full data set, using the θ you found; use the average CV error as its error estimate

B. Simple Linear Regression (30 points)

This exercise will introduce you to building and fitting linear regression models and some of the process behind it. We first examine a toy problem, focusing our efforts on fitting a linear model to a small dataset with three observations. Each observation consists of one predictor x_i and one response y_i for $i = 1, 2, 3$,

$$(\mathrm{x}, \mathrm{y}) = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$$

To be very concrete, let's set the values of the predictors and response.

$$(\mathrm{x}, \mathrm{y}) = \{(1,2), (2,2), (3,4)\}$$

There is no line of the form $\beta_0 + \beta_1 x = y$ that passes through all three observations, since the data are not collinear. Thus, our aim is to find the line that best fits these observations in the least-squares sense.

1. Follow the instructions below (10 points)
   1) Create two numpy arrays `X_train` and `y_train` with this data.
   2) Check the shapes of these arrays. They should be `(3,1)` and `(3,)`, respectively.
   3) Make a scatter plot of this data. (there will be only three points)

Linear regression is special among the models in machine learning because it can be solved explicitly. While most other models (and even some advanced versions of linear regression) must be solved iteratively, linear regression has a formula where you can simply plug in the data.

For the single predictor case it is:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1)$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (2)$$

Where $\bar{y}$ and $\bar{x}$ are the mean of the y values and the mean of the x values, respectively.
From the re-arranged second equation we can see that the best-fit line passes through $(\bar{x}, \bar{y})$, the center of mass of the data.
From any of the first equations, we can see that the slope of the line has to do with whether or not an x value that is above/below the center of mass is typically paired with a y value that is likewise above/below, or typically paired with one that is opposite.

2. Follow the instructions below (10 points)
    1) Complete the function `simple_linear_regression_fit` in Lab09_B.py. This function takes predictors and responses as its input, solves the best-fit line using equations 1 and 2, then returns β_0 and β_1.
    2) Find the best-fit line for the training data given in question 1 by using `simple_linear_regression_fit(X_train, y_train)`. Do the values of β_0 and β_1 seem reasonable to you?
    3) Plot the best-fit line together with the training data.

Now that we can concretely fit the training data from scratch, it's time to show how to implement simple linear regression with Python package, *scikit-learn*.
In scikit-learn, an *estimator* is a Python object that implements the methods `fit(X, y)` and `predict(T).` For simple linear regression, we will use the estimator, `LinearRegression`, in `sklearn.linear_model`.

3. Follow the instructions below (10 points)
    1) Find the best-fit line for the training data given in question 1 using `LinearRegression` estimator.
    2) Does the best-fit line founded using sklearn same as the one you found in question 2?
    3) Plot the best-fit line together with the training data.

C. Multiple and Polynomial Regression (35 points)

This exercise will explore a truly interesting dataset. This dataset, **epldata_final.csv**, were scraped by Shubham Maurya and record various facts about players in the English Premier League. The aim of this exercise is to find a model that predicts the players' market value (what the player could earn when hired by a new team).

`name`	Name of the player
`club`	Club of the player
`age`	Age of the player
`position`	The usual position on the pitch
`position_cat`	1 for attackers, 2 for midfielders, 3 for defenders, 4 for goalkeepers
`market_value`	As on transfermrkt.com on July 20th, 2017
`page_views`	Average daily Wikipedia page views from Sept. 1st, 2016 to May 1st, 2017
`fpl_value`	Value in Fantasy Premier League as on July 20th, 2017
`fpl_sel`	% of FPL players who have selected that player in their team
`fpl_points`	FPL points accumulated over the previous season
`region`	1 for England, 2 for EU, 3 for Americas, 4 for Rest of World
`nationality`	Player's nationality
`new_foreign`	Whether a new signing from a different league, for 2017/18 (till 20th July)
`age_cat`	A categorical version of the Age feature
`club_id`	A numerical version of the Club feature
`big_club`	Whether one of the Top 6 clubs
`new_signing`	Whether a new signing for 2017/18 (till 20th July)

In any machine learning problem, we want to make sure that the training and test data comes from the same distribution. This is especially important in this dataset because some regions are rather rare, it would be bad for the training data to entirely miss a region.

1. Follow the instructions below (15 points)
   1) Use the `train_test_split` function in `sklearn.model_selection` to split your data. Make sure that you are keeping equal representation of each region.
   2) Describe your approach and any of the issue that you encountered. How do you determine your test size? How do you make sure the training data and test data have appropriate representation of all features?

After exploring and splitting the data, we will now focus on fitting a model to the data and interpreting the model results. The model we'll use is

$$\text{market\_value} \approx \beta_0 + \beta_1 \text{fpl\_points} + \beta_2 \text{age} + \beta_3 \text{age}^2 + \beta_4 log_2(\text{page\_views}) + \beta_5 new\_signing + \beta_6 \text{big\_club} + \beta_7 \text{position\_cat}$$

2. Answer the questions below (20 points)
   1) How good is the overall model?
   2) Interpret the regression model. What is the meaning of the coefficient for:
      - age and age^2
      - $\log_2$(page_views)
      - big_club
   3) What should a player do in order to improve their market value? How many page views should a player go get to increase their market value by 10?
   4) How does the model perform on training and test data? Do you think the model overfits the training data?