# Machine Learning for Public Policy
# Spring 2019
# Mid-term Assignment

## Due: May 14, 5pm

**Instructions**:

- This is an individual assignment – please do not work in groups.
- This is open-book, open-notes, open-internet but no need for any programming to do any of the work.
- Please do not use any code to solve any of the problems here
- You should show your work instead of just giving me the answer.
- You can spend as much time as you want.
- Please submit the assignment on canvas as a pdf.

## (Short answers) [30 pts – 2 points each]

1. You're asked to predict the probability that the unemployment rate will go down next quarter for each of the neighborhoods in Chicago. Which model would you prefer to use?

   A. Logistic Regression
   B. Support Vector Machines

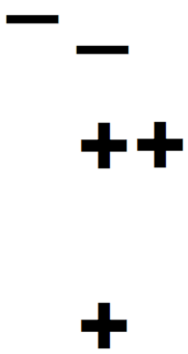
   Why?

2. Do you have to do anything special with the data or features for this problem with the model you chose in #1?

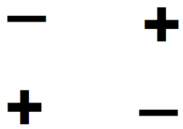
3. What is the training error (error on the training set) for a 1-NN classifier?

4. What is the Leave-one-out cross validation error for k nearest neighbor on the following data set? List any assumptions you may be making.

   A. For k=1
   B. For k=3



5. Which of the following classifiers are appropriate to use for the following data set? Why?



   A. Logistic Regression
   B. Decision Trees
   C. SVMs

6. You are being asked to build a model to predict which children in Chicago are at risk of Asthma attacks. You create 1000 features (all continuous) but you find out after exploring the data and talking to public health and medical experts that ~10 of them are useful and relevant for the prediction problem. Which of the classifiers below would you choose? And why?
   a. K-NN
   b. Decision Trees

7. Does Boosting give you a linear classifier? Why or why not?

8. Can boosting perfectly classify all the training examples for any given data set? Why or why not?

9. If you have a data set with 10 variables, each of them binary, with a binary target variable (label). You are asked to build a decision tree. If you didn't use a greedy approach and built all possible trees on this data set (without pruning or limiting the depth), how many trees would you build?

10. You are reading a paper for a new method to identify people at risk of pre-term and adverse births.. The reported accuracy is 89.4% and the precision at the top 10% is 56%. Are those numbers high enough to justify you replicating the method in your project (please explain your answer in 1-2 sentences)?
    a. Yes
    b. No
    c. Maybe

11. A Random Forest will always perform better than a decision tree on the same data set.

A. True

B. False

12-15. You need to build a model to predict re-entry into a social service program. A colleague suggests building a separate model for males and females while another colleague insists you just need to build one combined model.

12. When will separate models be more appropriate?

13. When will a combined model be more appropriate?

14. What are the pros and cons of each approach?

15. What is your opinion on how to proceed?

## Section B [55 pts]

### 1. Decision Trees [12 pts]

| Temperature | HomeInsulation | HomeSize | EnergyConsumption |
|---|---|---|---|
| Hot | Poor | Small | Low |
| Mild | Poor | Medium | High |
| Cool | Excellent | Large | Low |
| Hot | Excellent | Large | High |
| Hot | Excellent | Medium | Low |
| Mild | Poor | Small | High |
| Cool | Poor | Small | High |
| Cool | Excellent | Medium | Low |
| Cool | Excellent | Medium | High |
| Cool | Poor | Medium | High |

A. What will be the random baseline accuracy for this data set?

B. Calculate the entropy for the target variable, EnergyConsumption

C. Now calculate the Information Gain if you do a split on the feature "Home Insulation".

D. Using the data above, construct a two-level decision tree that can be used to predict Energy Consumption. Don't worry about overfitting or pruning. You can use a simple algorithm such as ID3 (using information gain as the splitting criterion).

## 2. Evaluation 1 [12 pts]

The table below shows the predictions of two classifiers, SVM and Logistic Regression for 10 examples. The classifiers are predicting the probability that the Label is 1.

| ID | Probability (assigned by SVM) | Probability (assigned by Logistic Regression) | True Label |
|---|---|---|---|
| 1 | 0.98 | 0.85 | 1 |
| 2 | 0.2 | 0.3 | 0 |
| 3 | 0.1 | 0.22 | 0 |
| 4 | 0.99 | 0.9 | 1 |
| 5 | 0.55 | 0.4 | 0 |
| 6 | 0.05 | 0.2 | 0 |
| 7 | 0.4 | 0.1 | 1 |
| 8 | 0.35 | 0.35 | 0 |
| 9 | 0.65 | 0.81 | 0 |
| 10 | 0.75 | 0.5 | 1 |

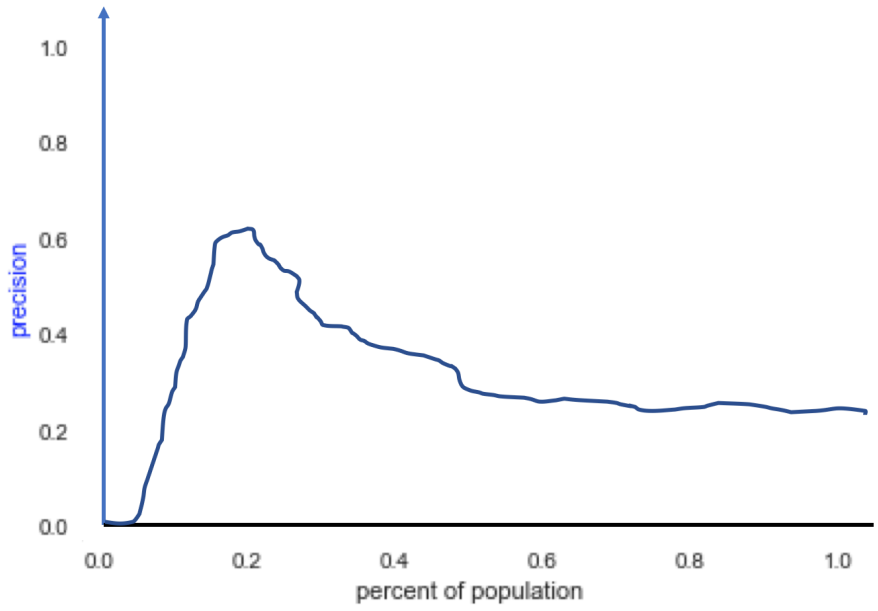
A. What is the accuracy of the SVM on this set? You will need to make some assumptions here. Be very explicit about your assumptions

B. Plot the precision recall curves for both classifiers based on these predictions.

C. Which classifier is better? (again, list the assumptions you're making)

**3. Evaluation 2 [12 pts]**

You just finished building and evaluating a model and got the following precision graph. You might notice that the precision is very low in the beginning of that graph.




A. How would you explain what's happening at the beginning of the graph to someone who's not a machine learning expert?

B. What could be the reason for that behavior?

C. What would you do to improve the performance of the classifier at the top 5%?

**4. Evaluation 3 [12 pts]**

You have trained three types of models on a training set and validated the results on a hold-out test set on a variety of metrics. The table below shows results from your trials.

| Model Type | parameters | baseline | Precision at 5% | Precision at 10% | Precision at 20% | AUC ROC |
|---|---|---|---|---|---|---|
| Decision Trees | 'max_depth': 1 | 0.34 | 0.36 | 0.32 | 0.40 | 0.50 |
| Decision Trees | 'max_depth': 20 | 0.34 | 0.27 | 0.25 | 0.27 | 0.45 |
| Decision Trees | 'max_depth': 100 | 0.34 | 0.30 | 0.33 | 0.31 | 0.49 |
| Logistic Regression | {'penalty': 'l1', 'C': 0.001} | 0.34 | 0.66 | 0.60 | 0.42 | 0.53 |
| Logistic Regression | {'penalty': 'l2', 'C': 0.001} | 0.34 | 0.73 | 0.53 | 0.45 | 0.57 |
| Logistic Regression | {'penalty': 'l1', 'C': 0.1} | 0.34 | 0.82 | 0.64 | 0.52 | 0.62 |
| Logistic Regression | {'penalty': 'l2', 'C': 0.1} | 0.34 | 0.82 | 0.65 | 0.53 | 0.63 |
| Logistic Regression | {'penalty': 'l1', 'C': 1} | 0.34 | 0.68 | 0.57 | 0.49 | 0.62 |
| Logistic Regression | {'penalty': 'l2', 'C': 1} | 0.34 | 0.77 | 0.68 | 0.54 | 0.62 |
| Logistic Regression | {'penalty': 'l1', 'C': 10} | 0.34 | 0.70 | 0.60 | 0.51 | 0.62 |
| Logistic Regression | {'penalty': 'l2', 'C': 10} | 0.34 | 0.75 | 0.65 | 0.51 | 0.62 |
| Random Forests | { 'n_estimators': 1000} | 0.34 | 0.59 | 0.55 | 0.47 | 0.61 |
| Random Forests | {'n_estimators': 10000} | 0.34 | 0.57 | 0.56 | 0.47 | 0.61 |

A. What can you say about the behavior of Logistic Regression as you vary the parameters?

B. Which specific model would you select to deploy (and why)? going forward if:

I) your goal was to prioritize 5% highest risk population to intervene with?

2)the resources available for interventions were yet to be determined?

**5. Communicating your results [12 pts]**

You have recently built a model that assigns a risk score to all students beginning 9th grade of not graduating high school within 4 years. You receive a call from the school administrator who asks you "According to your model, Jenny has a risk score of 50 (out of 100), but I know she is a bright student and has done well so far. Why has your model assigned her a score of 50 and not much lower?"

A. How would you explain this to the school administrator? Assume this administrator is a reasonable, intelligent person with extensive school administration experience and little or no background in statistics and machine learning.

B. Then suggest a different way that the administrator can confirm the accuracy of the predictive model you created.

**<u>Section C: Solving a New Problem</u> [10 pts]**

- Please do not use the internet or books for this question.

A critical part of most machine learning problems is integrating data from multiple sources about the same people, places, or businesses. For example, if you are working on predicting the risk of an individual going back to jail in order to inform preventative interventions, you might get data about that individual from the Department of Corrections, Department of Mental Health, Emergency Medical Services, and Homeless Shelters. Your first task is often to integrate that data and link records about the same people. This is known as record linkage (or matching) and is often done through exact matching or through "fuzzy" matching rules.

Now that you know how to do machine learning, your task is to come up with a machine learning solution for this problem of linking records that belong to the same person across different data sources.

You can assume that all data sources have some columns/fields in common, let's say first name, last name, date of birth, address, gender, and race.

A. How would you formulate this as a machine learning problem? (is it supervised learning or unsupervised learning? If it's the former, what's the label? What is each row in your training date? How would you get the training data?

B. What features would you create for this problem?

C. What models would you use?

D. What evaluation metric would you use? Be specific and justify your choioce.

E. Would you expect the machine learning solution to work better than exact matching or “fuzzy”/approximate matching rules? Why or why not?