6003 Assessment 2

- Time: 60 minutes
- Closed book
- Individual

There are 18 questions that cover these topics:

- Logistic Regression
- Model Evaluation
- SVMs
- Regularization
- Decision Tree
- Random Forest

We recommend skimming over all the questions and solving the ones that you are most confident about first.

Good luck!

Name:

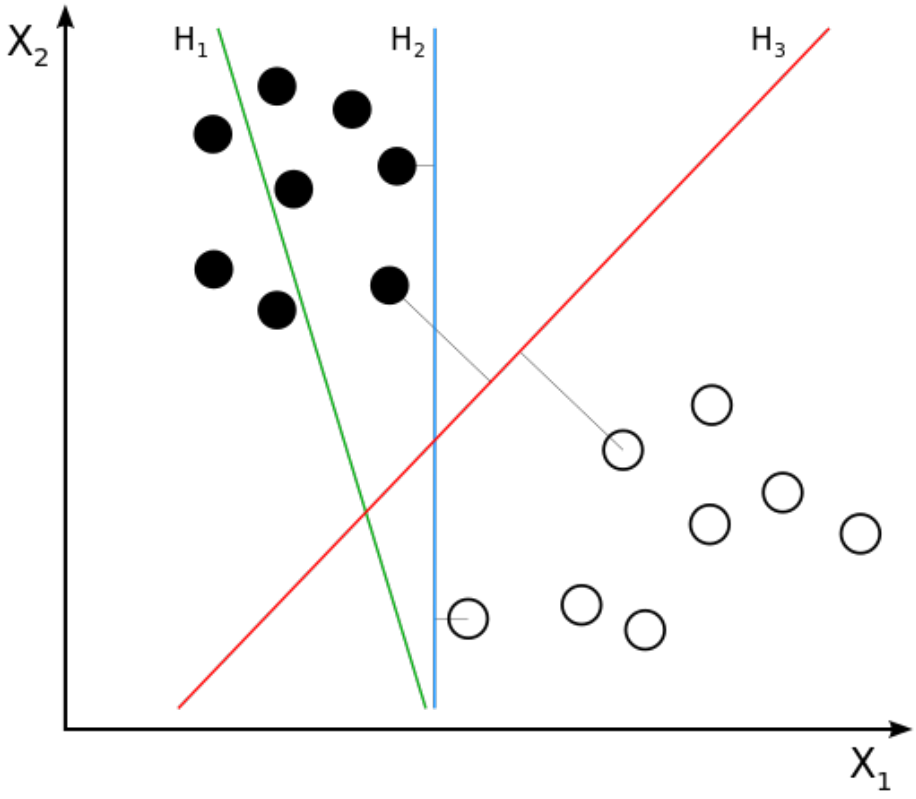
Date:

Assessment

1) You fit a linear regression to predict SAT score with many predictors, one of which is whether or not the student was homeschooled. `Beta_homeschool = -40`. How do you interpret the coefficient?

2) You fit a logistic regression to predict whether or not a student was admitted to a 4-year university. Again, `Beta_homeschool = -0.3`. How do you interpret the coefficient?

3) In the following image, label the decision boundary of the SVM(and explain why you chose it).



4) Give an example of a confusion matrix with precision $> 90\%$ and recall $< 10\%$ consisting of nonzero entries.

$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN}$$

Orient your confusion matrix like this:

```
         Predicted
         ---------
        | TP | FN |
Actual   ---------
        | FP | TN |
         ---------
```

5) Give an example of a confusion matrix with accuracy $> 90\%$, but both precision $< 10\%$ and recall $< 10\%$.

6) What are 2 benefits of using F_1 score during model selection/hyperparameter tuning, versus say using **accuracy** or a **confusion matrix**?

7) Say I'm building a Decision Tree Classifier on this dataset.

color	number	label
blue	1	0
blue	2	1
red	1	0
red	5	1

Splitting on what feature and value has the best information gain? Use your intuition rather than calculating all the entropy values.

8) Say I'm building a Decision Tree Regressor. What is the information gain of this split of the data?

   To calculate information gain, you only need the values of the label, so those are the values given. You are not given the feature values.

```
Split 1: 6, 5, 8, 8
Split 2: 5, 4, 2, 4, 4
```

   Hint: Use variance.

9) You build a Decision Tree and get these stats:

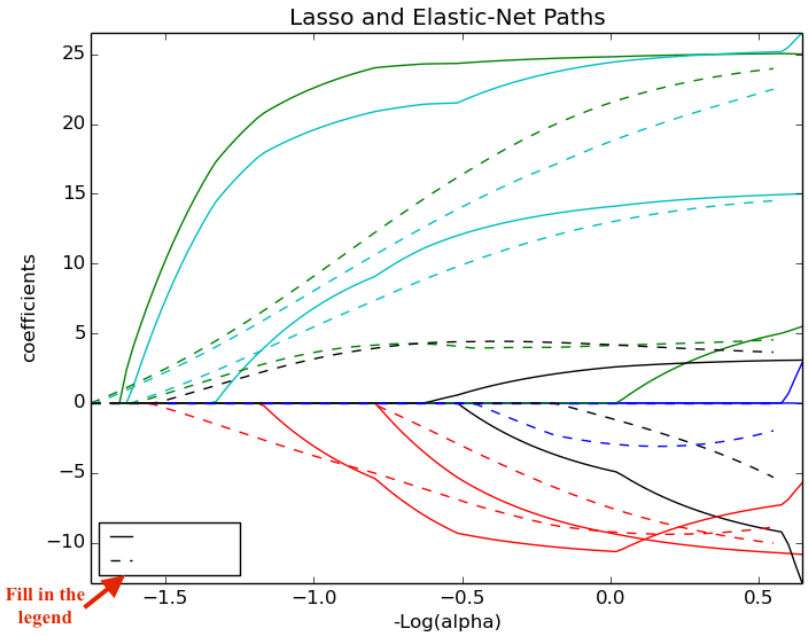
```
train set accuracy:  90%
train set precision: 92%
train set recall:    87%

test set accuracy:   60%
test set precision:  65%
test set recall:     52%
```

What's going on? What tactic(s) do we have to modify our Decision Tree to fix the issue?

10) How are the Decision Trees in Random Forests different from standard Decision Trees?

11) Why are we able to cross validate our Random Forest with our training set (OOB error)? I.e. Why doesn't this count as testing on my training set?

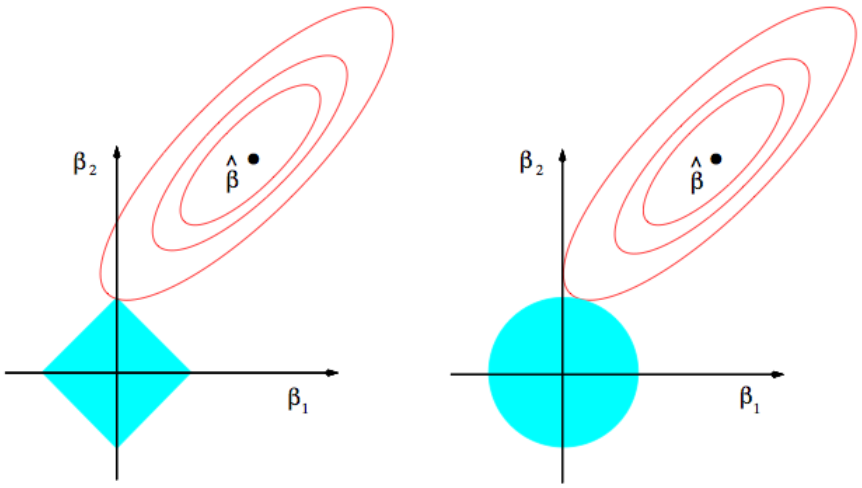
12) Label the following trace plot curves (dotted or solid) and explain your reasoning for each.



13) Name the tuning parameters for the following algorithms and describe what you'd expect to happen with bias and variance as you increase the tuning parameter. For the SVM classifier also describe how the decision boundary/margin changes as its hyperparameters change.

- Lasso / Ridge
- SVM

14) Label each of the following Lp space plots (with either Ridge, LASSO, or Elastic-Net) and explain your reasoning for each.



15) A logistic regression is fit to model whether or not a woman has an affair! Interpret the coefficients for 'yrs_married'. Observe the column of p-values, one for each predictor, and comment on any next steps you might take (more than 1 answer!).

Logit Regression Results

Dep. Variable:	affair	No. Observations:	6366
Model:	Logit	Df Residuals:	6357
Method:	MLE	Df Model:	8
Date:	Tue, 02 Dec 2014	Pseudo R-squ.:	0.1327
Time:	12:54:00	Log-Likelihood:	-3471.5
converged:	True	LL-Null:	-4002.5
		LLR p-value:	5.807e-224

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
Intercept	3.7257	0.299	12.470	0.000	3.140	4.311
occupation	0.1602	0.034	4.717	0.000	0.094	0.227
educ	-0.0392	0.015	-2.533	0.011	-0.070	-0.009
occupation_husb	0.0124	0.023	0.541	0.589	-0.033	0.057
rate_marriage	-0.7161	0.031	-22.784	0.000	-0.778	-0.655
age	-0.0605	0.010	-5.885	0.000	-0.081	-0.040
yrs_married	0.1100	0.011	10.054	0.000	0.089	0.131
children	-0.0042	0.032	-0.134	0.893	-0.066	0.058
religious	-0.3752	0.035	-10.792	0.000	-0.443	-0.307

Figure 1: Logistic Regression Output

16) What does it mean for an SVM to be a `maximum margin` classifier?

17) You have a bag full of marbles: 501 blue and 530 red to be exact.

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i)$$

a. What is the Gini impurity of your bag of marbles?

b. You split your bag of marbles into two bags! Bag1 has 1 blue and 30 Red. Bag2 has 500 blue and 500 red. What is the Gini impurity of Bag1 and Bag 2? What is the Information Gain in going from Bag –> Bag1 and Bag2?

c. Consider a different split where Bag1 has 100 blue and 400 red, and Bag2 has 401 blue and 130 red.
Again, use Gini impurity to compute the Information Gain.

d. Why is the gain much better in part (c), despite the purity of Bag1 in part (b)?

e. In the general classification tree context, for any given node, how is a particular split chosen?

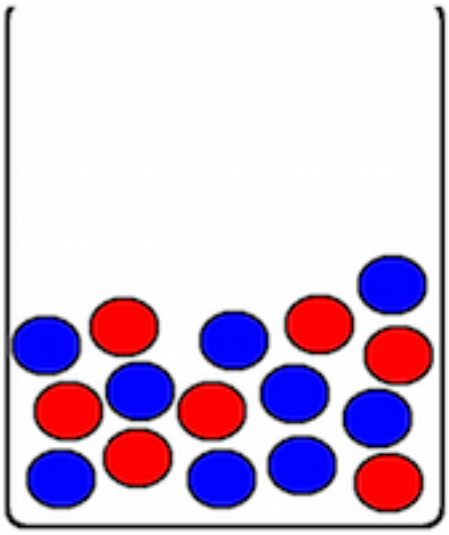


Figure 2: Marbles

18) Write a one sentence description of `precision` in plain English. Do the same for `recall`.