CSC 422/522: HW4

Total: points

Question 1 [5 Points] - Kaggle Intro

Kaggle `http://www.kaggle.com/` is a platform for predictive modeling competitions which are hosted by various agencies around the world. The recommended project for this course is one such competition. This exercise aims at introducing students to the Kaggle environment.

This is the easiest question in the entire course; please don't make a mistake. Complete the following task.

- Go to `http://www.kaggle.com/`
- If you are not a member already, signup.
- Go to the Database in Moodle and enter your unity ID, Kaggle user name and Kaggle userID.

Question 2 [25 Points] - Evaluation of Classifiers

For this exercise, you will use the getting started competition Titanic: Machine Learning from Disaster `http://www.kaggle.com/c/titanic-gettingStarted`. You may use Matlab, R or Weka. Note there is a tutorial on this contest posted on the course website. Submissions to Kaggle are scored on the basis of classification accuracy, which is the proportion of test cases that are correctly classified. The data can be found here `http://www.kaggle.com/c/titanic-gettingStarted/data` and forums with a lot of helpful discussion can be found here `http://www.kaggle.com/c/titanic-gettingStarted/forums`. Also note that you only get two submissions to Kaggle per day, so please submit early and often to ensure you can complete this assignment. Make sure your submissions match the format provided in the tutorial to avoid wasting submission.

Complete the following tasks.

1. All machine learning models have parameters that can be tuned. These parameters can either have a positive or negative effect on the performance of the algorithm. Cross-validation (CV) is one approach to identify a good model. For each of the algorithms below, identify the set of parameters that give the best predictive performance in terms of classification accuracy. You are free to do this in whatever way you wish, but in R you might consider using the *train* function in the *caret* package to help you with

this task. This reference may be helpful `http://caret.r-forge.r-project.org/training.html`

- A decision tree (ID3, Gini, C4.5, C5 or any other single decision tree)
- A boosted tree. Use Statistics Toolbox -> Functions -> Nonparametric Supervised Learning -> Ensemble Methods in Matlab or the *gbm* package in R. In R, the tuning parameters for this function are `interaction.depth` (how big each tree is), `n.trees` (how many trees to grow), `shrinkage` (a parameter to control overfitting). Weka has ensemble models as well.

Your answers must include at least 5 different sets of parameters that you tried. State why you changed the parameters for each set. Report the cross-validation scores for each of the 5 sets for each approach. Use k=5 for a k-way cross-validation. Note you do not have to report the Kaggle leaderboard score for each.

2. Pick two other classification algorithms not covered in class and tune them using cross-validation. For an incomplete list of classification algorithms in R, take a look at this page `http://caret.r-forge.r-project.org/modelList.html`. Compare the performance of the two new models against the decision tree with tuning parameters and boosted tree from the previous trials. State which model performed the best on the basis of CV scores.

3. Submit the model with the best cross-validation performance to Kaggle and report your score. Was this close to your CV estimate of the model's classification accuracy?

Question 3[15 Points] - Probabilistic Models

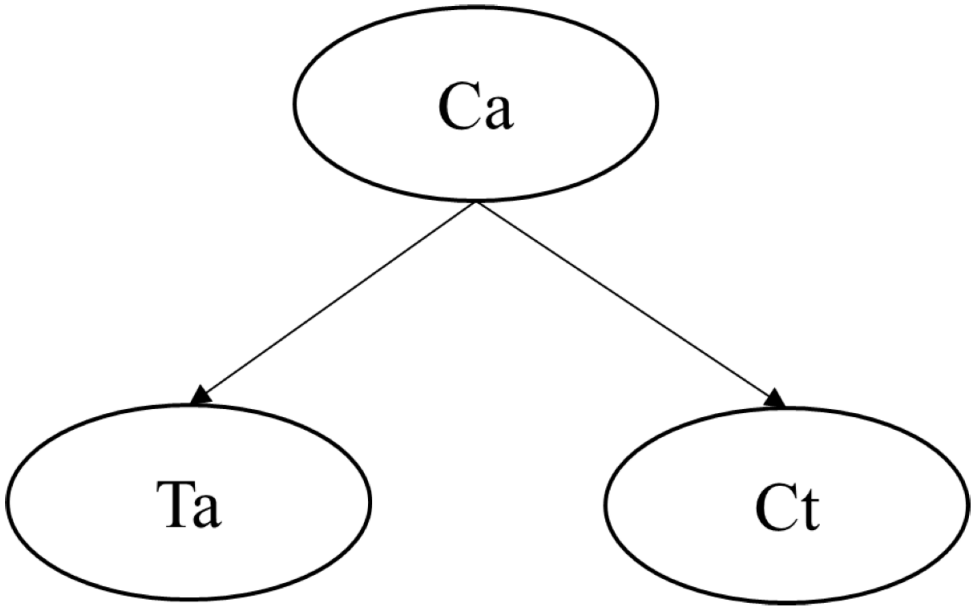
Consider the following Bayesian Network structure for a problem in a Medical domain.

The full joint probability distribution corresponding to the network is given in the table below.

	Ta		¬ Ta	
	Ct	¬Ct	Ct	¬Ct
Ca	0.108	0.012	0.072	0.008
¬ Ca	0.016	0.064	0.144	0.576

Based on the given information, calculate the following. Show your computations.

1. P(Ta)
2. P(Ca)
3. P(Ta | Ca)
4. P(Ca | Ta $\vee$ Ct)


Ca
Ta
Ct