KAGGLE ASSIGNMENT

URL: http://www.kaggle.com/c/titanic-gettingStarted

Overview of the assignment

[NOTE: You are no longer required to link your accounts. Some students had reasonable objections to the requirement to make their profile public. Questions related to this requirement are still included for reference, but you are not required to answer them, and no points can be earned from answering them.]

In this assignment, you will be participating in an **active** Kaggle competition. You can see the active competitions by filtering the list with the controls at the left.

This project is VERY open-ended. You may use whatever combination of techniques and tools you think are appropriate to generate a solution.

You will be graded by a simple peer review. You will turn in a brief writeup (a few sentences each) of your problem, your solution, your evaluation, and some improvement. The reviewers are not expected to judge the technical efficacy of your approach, but to make sure that the descriptions are clear and sensible.

Clarity is more important than technical depth in this exercise -- you are trying to briefly explain your approach and how well it worked. Think of it as an elevator conversation rather than a full report.

You will need to submit your writeup by **6/17** in order to allow time for a peer review in the final week. This does not give you much time, so don't bite off more than you can chew! The idea is to get a good, simple, comprehensible solution rather than try to impress people with how smart you are. Start small, and improve your solution incrementally -- don't try to save the world, boil the ocean, or any other hyperbolic metaphors.

If you are unsure where to begin, consider working on the Titanic Competition. There is some great help available for that competition in the kaggle forums, and we discuss it in the lectures as well.

If you've already completed kaggle competitions, all you need to is briefly write up you solution and you're done! But of course, it may be fun to try another competition....

HAVE FUN!!!

PEER EVALUATION

For your selected competition, write a few sentences describing the competition problem as you interpreted it. You want your writeup to be self-contained so your peer-reviewer does not need to go to Kaggle to study the competition description. Clarity is more important than detail. What's the overall goal? What does the data look like? How will the answers be evaluated?

Example: "The task is to predict whether a given passenger survived the sinking of the Titanic based on various attributes including age, location of the passenger's cabin on the ship, family members, the fare they paid, and other information. Solutions are evaluated by comparing the percentage of correct answers on a test dataset."

The overall task, as I understood from the Kaggle description, is to predict the number of people who survived in the Titanic disaster, using machine learning techniques. 2 files are used for this purpose - a training file that contains passenger information such as name, age, family members, class, the fare for the ride, among other information. This can be used to construct and train a model. This model can then be run over the test file to predict whether or not the passengers survived. The test model contains the same information as the training file but without the prediction column "survived".

The data contains information about the passenger, his siblings, the class he was in the ship, the cabin, where he embarked and how much fare he paid. There are 4 numerical fields: age, number of sibling/spouses, number of parents/children, and fare. The categorical variables are survived (0/1), sex(male/female), pclass(1/2/3), and embarked port (S/C/Q). The string variables are name, and ticket number.

The predictions are evaluated by comparing the percentage of correct answers on the test data.

Write a few sentences describing how you approached the problem. What techniques did you use?

Example: "I split the data by gender and handled each class separately. For the females, I trivially classified all of them as "survived." For the males, I trained a random forest as a classifier. I ignored the pclass atribute that indicated the location of the passenger's cabin because I didn't think it was relevant."

I looked at the training data and saw a few columns that were not relevant to the prediction. So I chose to disregard them from the training model.
Specifically, I disregarded the columns - name, sibsp, parch, ticket, cabin, and embarked.

Next, I constructed a model using Support Vector Machines (SVM). I chose SVM because its simplicity combined with state of the art performance on many learning problems (classification, regression, and novelty detection) has contributed to its growing popularity. The SVM model maps the input data into a high-dimensional feature space defined by a kernel function.I tuned the model was done by manipulating 2 variables – gamma and cost, used by the kernel function.

My approach is described as follows:

I took the training data and created a SVM training model using the svm() function, with survived as the output variable and pclass, sex, age, and fare as the input variables, and using the radial kernel option, with gamma = 0.1 and cost = 10.

I then tuned the SVM model using the tune.svm() function that gives the best value for gamma and cost. I used gamma over the range 10^-6 to 10^-3. and cost varying from 10 to 100. I used a 10-fold cross validation set.

Once the tuning was done, it gave me the best values for gamma and cost. I now ran the svm() function again over the train set with these best values of gamma and cost.

Finally, I applied this training model over the test set using the predict() function and got my predictions.

Write a few sentences describing how you implemented your approach. What languages and libraries did you use? What challenges did you run into?

Example: "I partitioned the data by gender manually using Excel. I used Weka to build the random forest."

I used R language as it provides numerous packages to do predictive modeling. For svm, I downloaded the library e1071 package. At first, I used the survived column vs all columns for the model. I ran into blank values in the age and cabin columns (reported as NA), and factor issues as some columns like sex are factor columns (male/female). This gave me the idea that I could use the median age for blanks in the age column, I ignored the cabin, and used only the numeric columns and converted sex to numeric. I was then able to generate the model successfully. I then filled blank values in a similar manner for the test data, and did the prediction.

Write a few sentences assessing your approach. Did it work? What do you think the problems were?

Example: "My approach did not work so well, achieving a score of 0.65. This is less than the sample solution. I suspect I should not have ignored the pclass attribute."

I thought since I was using SVM, a sophisticated R prediction package, I would have better results. But it turns out that I had a prediction score of 076555 which certainly fell below my expectations.of SVM. Maybe I could have considered some of the other variables such as cabin or sibsp, and it could have helped improve the accuracy.

The prediction did work though, and I was satisfied that I had put in a sincere effort to learn SVM to get this far.

Write a few sentences describing how you improved on your solution, and whether or not it worked.

Example: "I included the pclass attribute and ignored the ticket number attribute. However, my score was actually lower the second time!"

In an effort to improve prediction accuracy, I included the cabin column as a numeric value and added it to the other svm input variables. I then used the svm training model and did the prediction again, but it was still the same ie. a score of 0.76555 on Kaggle.