MACHINE LEARNING
BARUCH COLLEGE
SPRING 2015
MIGUEL A. CASTRO
T.A.: LIQUN ZHU

Homework Assignment 4

1. You've constructed a model to classify profitable vs. unprofitable trades on price series data that show an *a priori* (pre-classification) ratio of profitable vs. unprofitable opportunities given by $\frac{\#\,Profitable\,Trades}{\#\,Unprofitable\,Trades} \equiv H$. Show that, in order to be profitable, your classifier must lie in the region bounded from below by the line $y = \left(\frac{1}{2H}\frac{L}{P}\right)x + \frac{1}{2}$ in the ROC plane. More specifically, in a liquidity-providing HFT strategy, the loss-to-profit ratio of a trade is typically given by: $\frac{L}{P} = \frac{1+2c}{1-2c}$, where c is the transaction cost in units of cents, so that $\frac{L}{P} \cong 3$; while the ratio of *a priori* profitable vs. unprofitable trades is of the order of $1/2$. Plot the region in ROC space corresponding to the viable classifiers for this strategy.

2. You're searching for Association Rules in a transaction database with 1,000,000 transactions, but this time your algorithm is using the Dependency Framework. You've decided to use Minterm Support of $s = 5$ at the $p = 80\%$ level, and a χ^2 Significance Level of $\alpha = 0.05$ (i.e., a 95% Confidence Level). You're considering Itemsets X and Y and have measured the following occurrences in the database: Number of occurrences of X is 500,000; Y occurs 400,000 times. (a) Assuming that X and Y occur together 200,000 times, what is the Minterm Support? Does it clear the Minterm Support threshold? (b) Compute the χ^2 Statistic, and notice that it does not clear the Significance Level. (c) Based on your answers to (a) and (b) is this an Interesting Association Rule according to the Dependency Framework? Discuss whether this is a good idea and contrast with your answer to 5.(c) in Homework Assignment 3 from last week. Now suppose that X and Y occur together only 2,000 times. (d) What is the Minterm Support? Does it clear the Minterm Support Threshold? (e) Compute the χ^2 Statistic, and notice that it does clear the Significance Level (alternatively you could look up this value in a table for a χ^2 Statistic with degree of freedom = 1). (f) Based on the answers to (d) and (e), and assuming the Dependency Strength is high enough, is this an Interesting Association Rule according to the Dependency Framework? Discuss whether this is a good idea and contrast with your answer to 5.(f) from Homework Assignment 3.

3. What is the Region of Interest and how is it generated? In the Dependency Framework, why is the Dependency Strength Coefficient U not used for generating the Region of Interest? Given the same database $\mathcal{D}$, would you expect the Region of Interest to be smaller or larger for the Dependency Framework as compared with the Support-Confidence Framework? Discuss.

4. You're searching for Association Rules in a very large transactional database, and you've decided to use sampling to reduce the computation time. However, you do not want the probability that your support error exceed 0.1% to be more than 0.01. What should your sample size be?

5. Within the classical Support-Confidence Framework for finding Association Rules, prove that Support is Internally Closed. If you find an Itemset that is not supported, how can we use this property (along with Closure Complementarity) to prune the search space? Is Support a Type I or a Type II Property? Why?

6. The Dependency Framework uses the significance level α to find Interesting Association Rules, but ignores β. Suppose that you have obtained a set of Dependency-Framework Association Rules using a given $\alpha \ll 1$, and that you have estimated β using a non-central χ^2 distribution. Suppose, further, that you have estimated the Association Rules' True Positive Rate as $1 - \alpha$, and that the loss incurred by an incorrectly assumed actionable for your Dependency Framework is *L*, while the benefit obtained by a correct actionable is *P*. What would be the minimum power for your Framework to be profitable?

7. An ANN with a Tapped-Delay Line with d delays is used as a predictor for a time series y_t. Assuming that the network has a sufficient number of hidden nodes, and that you've adequately regularized the network and used techniques such as early stopping to avoid overfitting, what would be the advantage of using such a network to model the series y_t over using an Autoregression of the form

$$y_t = \sum_{k=1}^{d} \varphi_k y_{t-k} \,,$$

where the φ_i are obtained through linear regression,?