DATA MINING NOTES

BRENDAN WHITAKER

2010 *Mathematics Subject Classification.* 12-XX

Abstract. A comprehensive set of notes for Data Mining course, taken SP18 at The Ohio State University.

Contents

Part 1

CHAPTER 1

Introduction

Wednesday, January 10th

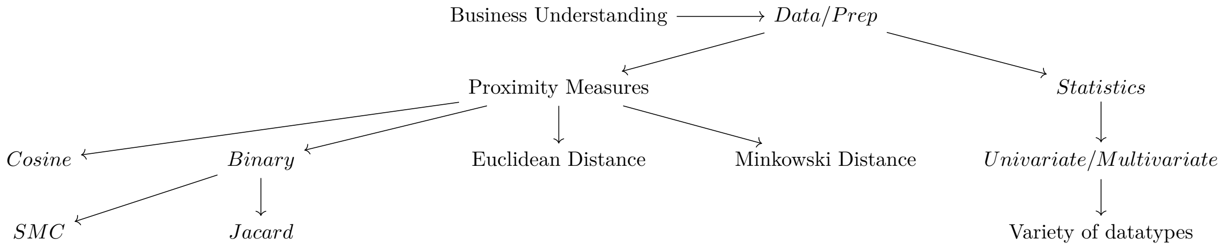
Note *class* = *label* = *category*. These are the dependent variables, the stuff that our work is determining.

Ratio variables are your typical real-valued numbers, i.e. zero is meaningful. Interval variables have 0 as just another possible value, it is not the additive identity in this case.

Monday, January 22nd

Review of stuff up to now.

$$\begin{tikzcd}
Data \arrow[rr] \arrow[dr, dash] & & Insights \arrow[dl, dash] \\
& CRISP-DM &
\end{tikzcd}$$



$$\begin{tikzcd}
\text{Exploratory Data Analysis} \arrow[r] & Visualization
\end{tikzcd}$$

Imputation: replacing missing data with the mean is one way we could do it. We could also use regression imputation to estimate the value of the missing variable using the data from other variables. You could also use a random value.

Monday, January 29th

PCA - principal component analysis.

Monday, Febrary 5th We have:

$$P(+|x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2}}.$$

Okay so last time, we talked about how you might compute a ROC curve. Recall that a ROC curve plots false positive rate (FPR) on the x-axis and true positive rate (TPR) on the y-axis. So we are going to set our $t = 0$.

1.1. Exam 1

Need to know proximity measures, types of data (ratio).

- Won't ask to compute covariance, that's mean.
- Know bayes theorem for bayes classifiers.
- PROXIMITY MEASURES, INFER IF WE'RE TALKING ABOUT A SIMILARITY VS DISSIMILARITY.
- know minkowski distance.
- don't know mahalaboni distance.
- simple matching and jaccard coefficients, and cosine.
- you use jacard when data is SPARSE, otherwise SMC.
- Make a reference sheet to use on exam.

- You'd use cosine when it is no longer binary data, but it can be used for anything.
- SMC and Jaccard are for BINARY ONLY.
- Know that the end of boxplots are 90 and 10.
- Interpret this scatterplot or boxplot on exam. Any of the data vis types.
- What is the difference between noise and outliers?
- Principal component analysis is going to be fair game, or part of the calculation, how do you do it, what does it mean, what are you trying to accomplish.
- how many principal components do you keep given some type of data loss requirement.
- Difference between dimensionality reduction, and feature subset selection.
- Dimensionality reduction creates new features, no longer interpretable, the values are meaningless, each of the new dimensions are combinations/transformations of the old ones.
- Feature subset selection just removes a subset of the n dimensions and uses those, which preserves the meaning of the dimensions/features.
- When you run PCA, you aren't doing anything related to the class variable.
- But Feature subset selection on the other hand, is usually done relative to a classifier.

Now let's talk about classification.

- What is objective of classification: create a generalizeable, predictive model. It has to GENERALIZE!
- As you train your model and it becomes more complex, training error keeps dropping, but test error find a local min and then goes up again.
- Producing a classification label is a two step process.

Example 1.1. Suppose we're doing a binary classification.
Step 1:
For a given record x, define:

$$\begin{aligned} P(+|x) &= 0.64 \\ P(-|x) &= 0.36 \end{aligned} \tag{1.1}$$

For logistic regression, we have:

$$P(+|x) = \frac{1}{1 + e^{-*\beta_0 + \beta_1 x)}}.$$

So x enters our model and get the the above posterior out of it.

Step 2:

We set some threshold t and say $x \in +$ if the posterior is $\geq t$ and in $-$ otherwise. And 0.5 is our default t value. You get an ROC curve only by varying t. Note the confusion matrix is dependent on the value of t.

- A ROC curve has FPR on x axis and TPR on y axis.
- When we increase t, precision generally goes up because we expect the model to work better than random.
- Recall precision is TP/(TP + FP), it is the top left box over the sum of the left half of the matrix.
- You should know DECISION TREES, NAIVE BAYES, kNN. Ensembles just a little bit. Rule based classifier.
- **What is reduced-error pruning?** Before we talk about that, normal pruning is just computing error for all nodes and removing nodes that don't give you lift (reduction of error rate). Reduced error pruning using validation data, you compute error rates for each node on the TEST DATA instead, then compute lift and decide to prune any nodes that don't give you a reduction in error.
- for notes sheet just print a shitton of these pages.

1.2. Final Notes

Study maximal and closed itemsets, there will be a question on the final on this.
Monday, March 19th
Wednesday, April 11th
Projects due the 18th of April. Exam is on the stuff from the second half of the course.

1.3. Scratch work

$$f(x) = \begin{cases} \frac{1}{k+1} & \text{if } x = \frac{1}{k}, k \in \mathbb{N} \\ x & \text{otherwise} \end{cases}. \tag{1.2}$$

1.4. Exam 2

Know how to do hierarchical clustering with dendrograms from the book exercises. Make sure you know the difference between similarity vs dissimilarity algorithm.

DBSCAN doesn't deal well with differing densities. It has global parameters.