COMP 330: Intro to Supervised Learning

Chris Jermaine and Kia Teymourian

Rice University

Phew!!!

Done with most math intense portion of class

"I appreciate that it gave me a realistic view of what data science is, even if it was a little dry."

But keep in mind

- Data science is a broad field!!
- Includes databases
- Includes big data systems
- Includes optimization
- Includes machine learning
- Includes probability and statistics

Possible to be a data scientist without hardcore math

That Being Said...

Almost no math today!!

"Supervised" Learning

One of the most fundamental problems in data science

- Given a bunch of (x_i, y_i) pairs
- Goal: learn how to predict value of y from x
- Called "supervised" because have examples of correct labeling

Problem Examples

From my own research:

- Given a text EMR, label "breast cancer" or not
- Given a document (email) in a court case, figure which subjects relevant to
- Given information about a patient surgery, predict death
- Given head trauma patient info, predict ICP crisis
- Given an set of surgical vital signs, label "good surgery" or not
- Many others!

Two Most Common Examples of SL

Classification and regression

Classification:

- Outcome to predict is in $\{+1, -1\}$ ("yes" or "no")
- Ex: Given a text EMR, label "breast cancer" or not

Regression:

- Outcome to predict is a real number
- Ex: Given an ad, predict number of clickthrus per hour

What Models Are Used?

Many!

- We will cover a number of them
- Simplest, most common: linear regression. From x_i, predict y_i as:

$$\sum_j x_{i,j} r_j$$

- $\langle r_1, r_2, \ldots, r_m \rangle$ are called regression coefficients
- Other common ones: kNN, support vector machines

Measuring Classification Accuracy

Simplest: % correct

- Pros and cons?

Measuring Classification Accuracy

Simplest: % correct

- Pros and cons?

False positive and false negative

- False positive: % of those we say are “yes” that are not really “yes”
- False negative: % of those we say are “no” that are not really “no”

Almost equivalent: Recall and precision

- Recall: % of those that are really “yes” that we say are “yes”
- Precision: % of those that we say are “yes” that are really “yes”
- Pros and cons?

Measuring Classification Accuracy

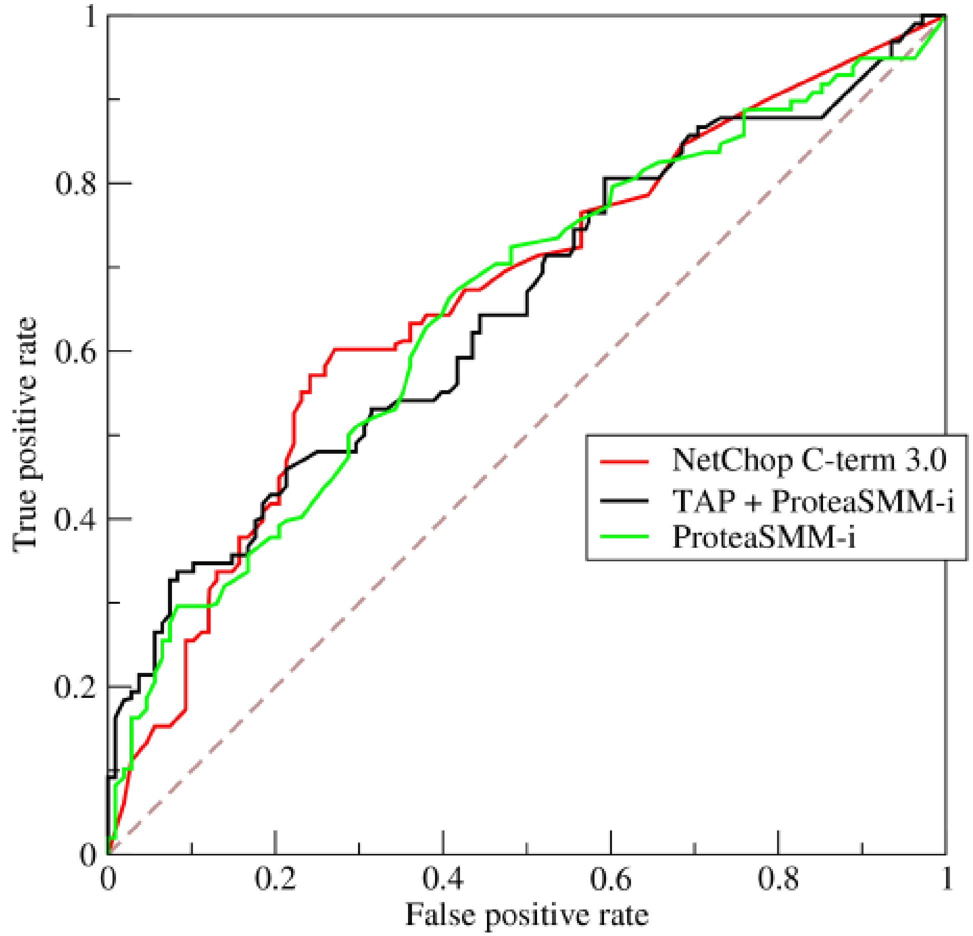
F_1

- Puts recall and precision into single number

$$F_1 = \frac{2 \times \text{ precision } \times \text{ recall}}{\text{precision } + \text{ recall}}$$

- Pros and cons?

AUC ROC



- ROC = "Receiver operating characteristic"
- AUC = "Area under curve"
- Gives single number from 0.5 to 1.0

- Less than 0.5 means “actively bad”
- Pros and cons?

Measuring Regression Accuracy

View the list of prediction errors as a vector

Can have many loss functions, corresponding to norms

Given a vector of errors $\langle \epsilon_1, \epsilon_2, ..., \epsilon_n \rangle$, l_p norm defined as:

$$\left(\sum_{i=1}^{n} |\epsilon_i|^p \right)^{1/p}$$

Common loss functions correspond to various norms:

- l_1 corresponds to mean absolute error
- l_2 to mean squared error/least squares
- l_∞ corresponds to minimax

Feature Selection

Lots of focus in supervised learning on models

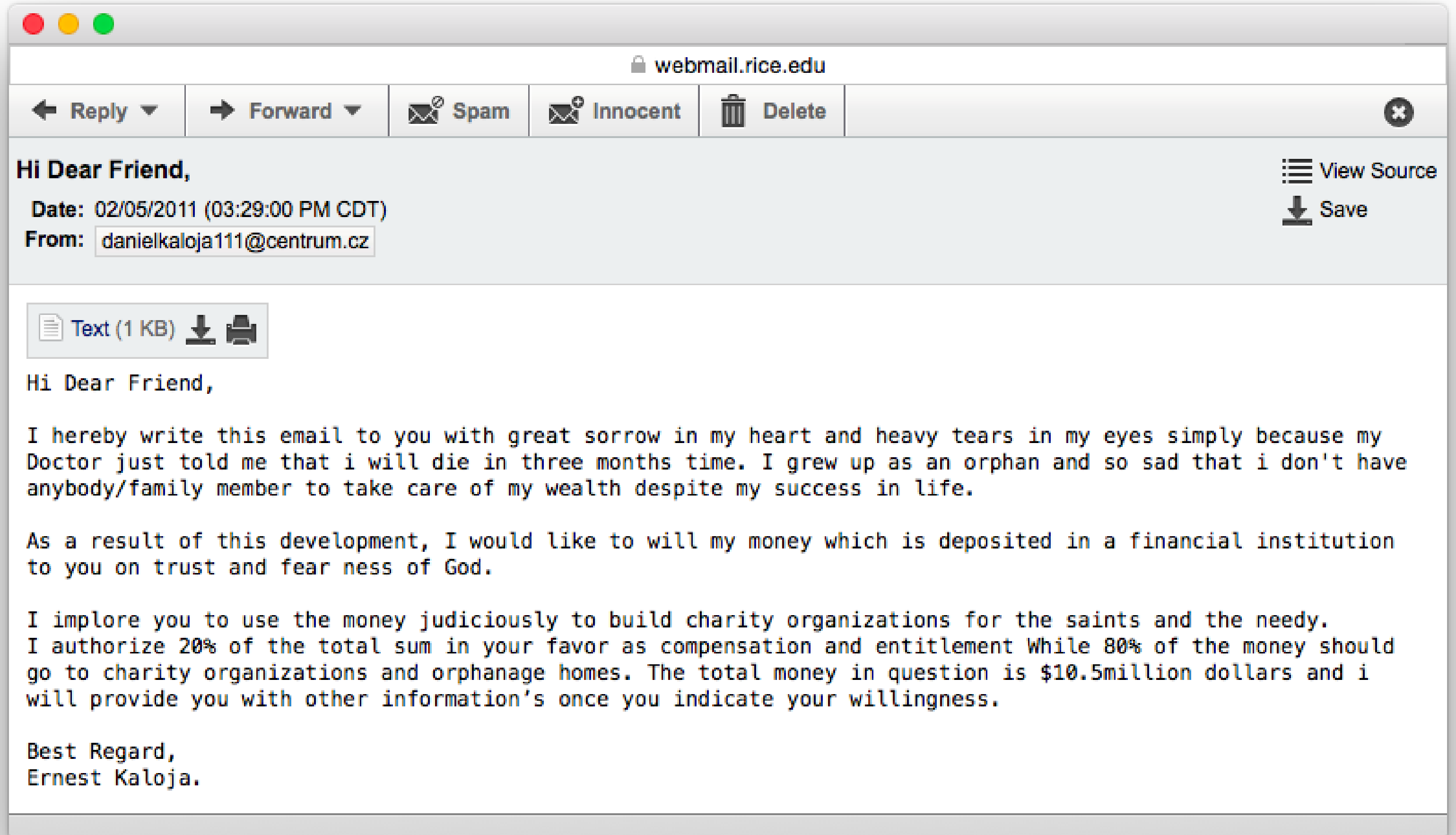
- Linear regression, SVM, kNN, etc.

Almost always less important than feature engineering

- That is, most simple models accept $x_i = \langle x_{i,1}, x_{i,2}, \ldots, x_{i,m} \rangle$
- Do not accept your raw data!
- How you “vectorize” is often the most important question!

Let’s consider feature engineering thru an example...

Example Feature Selection



webmail.rice.edu

Reply | Forward | Spam | Innocent | Delete

Hi Dear Friend,

Date: 02/05/2011 (03:29:00 PM CDT)
From: danielkaloja111@centrum.cz

View Source
Save

Text (1 KB)

Hi Dear Friend,

I hereby write this email to you with great sorrow in my heart and heavy tears in my eyes simply because my Doctor just told me that i will die in three months time. I grew up as an orphan and so sad that i don't have anybody/family member to take care of my wealth despite my success in life.

As a result of this development, I would like to will my money which is deposited in a financial institution to you on trust and fear ness of God.

I implore you to use the money judiciously to build charity organizations for the saints and the needy.
I authorize 20% of the total sum in your favor as compensation and entitlement While 80% of the money should go to charity organizations and orphanage homes. The total money in question is $10.5million dollars and i will provide you with other information's once you indicate your willingness.

Best Regard,
Ernest Kaloja.

“Bag of Words”

Might build a dictionary

- That is, map from each of m unique words in corpus
- To a number from $\{1 \ldots m\}$
- Then, each email is a vector $\langle 1, 0, 2, 1, 0, 0, \ldots \rangle$
- jth entry is num occurrences of word j
- Problems?

TF-IDF

"Term Frequency"

- Defined as:

$$TF = \frac{\text{num occurs of word in doc}}{\text{num words in doc}}$$

"Inverse Document Frequency"

- Defined as:

$$IDF = \log \frac{\text{num of docs having the word}}{\text{num of docs}}$$

TD-IDF defined as $TF \times IDF$

N-Grams

Words in this doc might not be suspicious

Might be how they are put together

- "great sorrow"
- "heavy tears"
- "financial institution"
- "fear ness"

Idea: also include all 2-grams, 3-grams, 4-grams, etc. as features

What Else?

Country of sender

Number of words in email

Time of day sent

Was the email sent previously?

Recipient list disclosed?

Supervised Learning Methodology

Important to divide available data into

- Training–used to learn model
- Validation–used to see if model useful
- Testing–used to evaluate useful models

Don't touch testing until ready to eval

- Evaluation on testing must be very last step!
- Why?

How To Perform Testing

One-off

- Apply validated model(s), get results
- Problems?

k-Fold Cross-Validation

- Break into k random subsets ("folds")

```
For i = 1 to k do:
  Train on all folds except i;
  Eval learned model on fold i;
Report average results;
```

Questions?