COMP 543: Tools & Models for Data Science

Intro to Supervised Learning

Chris Jermaine & Risa Myers

Rice University

“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 research done at Rice:
  - Given a text clinical note, label “breast cancer” or not
  - Given a document (email) in a court case, figure which subjects pertain
  - Given information about a patient surgery, predict death
  - Given head trauma patient info, predict intracranial pressure crisis
  - Given an set of surgical vital signs, label “good surgery” or not
  - Predict location and damage from a hurricane
  - Many others!

Two Most Common Examples of Supervised Learning

- Classification and regression
- Classification:
  - Outcome to predict is in $\{+1, -1\}$ (“yes” or “no”)
  - Ex: Given a text clinical note, label it as “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:

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

  - Where $x_{i,j}$ is a matrix of rows of feature values, one for each data point
  - $\langle r_1, r_2, \ldots, r_m \rangle$ are called regression coefficients
    - The relative magnitude of the regression coefficient tells you how important each feature is
  - Other common ones: kNN (A4), support vector machines

What do we Mean by Features?

- Clickthru
  - Location
  - Font
  - Size
  - Color
  - ...
- Patient
  - Age
  - History of Smoking
  - History of Diabetes
  - Blood pressure
  - Weight
  - ...

Measuring Classification Accuracy: Confusion Matrix

- Tabular representation of results

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN
Total		P	N

Measuring Classification Accuracy: % Correct

- Simplest: % correct
- $\frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} = 0.94$

? Pros and cons?

		Actual	
		Positive	Negative
Predicted	Positive	10	5
	Negative	2	95
Total		12	100

Measuring Classification Accuracy: % Correct

- Simplest: % correct
  - Pros
    - Easy to interpret
    - Easy to compute
    - Easy to compare
  - Cons
    - Certain classes may be more important than others
    - E.g. Male breast cancer
    - Can achieve high accuracy just by saying “No”
    - In this case, better to find all the cases and some that aren’t

Measuring Classification Accuracy: FP & FN

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN
Total		P	N

		Actual	
		Positive	Negative
Predicted	Positive	10	5
	Negative	2	95
Total		12	100

- False positive: % of those we say are “yes” that are not really “yes”

$$\frac{FP}{N} = \frac{FP}{FP+TN} = \frac{5}{5+95}$$

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

$$\frac{FN}{P} = \frac{FN}{FN+TP} = \frac{2}{2+10}$$

- Simple concept, easy to get wrong
- Male breast cancer: 50% FP rate is okay: 3 real cases + 3 extra cases
- Alert fatigue: 72 - 99% of alerts are not actual causes for alert [1]

[1] Sendelbach S, Funk M. Alarm fatigue: a patient safety concern. AACN advanced critical care. 2013;24(4):378-86.

Measuring Classification Accuracy: Recall & Precision

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN
Total		P	N

		Actual	
		Positive	Negative
Predicted	Positive	10	5
	Negative	2	95
Total		12	100

- Recall (Sensitivity): % of those that are really “yes” that we say are “yes”

$$\frac{TP}{P} = \frac{TP}{TP+FN} = \frac{10}{10+2}$$

- Precision: % of those that we say are “yes” that are really “yes”

$$\frac{TP}{TP+FP} = \frac{10}{10+5}$$

? Pros and cons?

Measuring Classification Accuracy: Recall & Precision

- Recall and precision
  - Pros
    - More information than accuracy
    - Tolerance for TP and FN can be different, based on the situation
  - Cons
    - Can be confusing

Measuring Classification Accuracy F1-score

- 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?

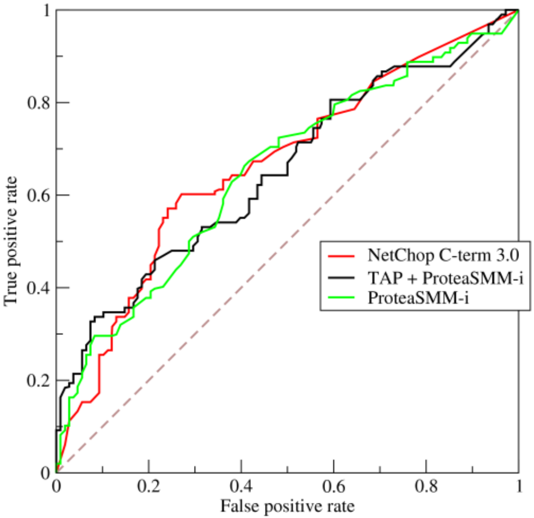
Measuring Classification Accuracy F1-score

- F_1
  - Puts recall and precision into single number

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

  - Pros
    - Reasonable way to combine precision and recall
  - Cons
    - Somewhat arbitrary
    - Could use F_2, F_3, etc.

AUC ROC

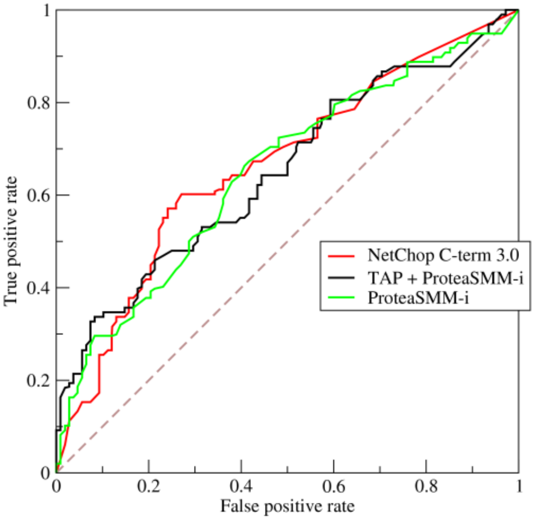


- ROC = “Receiver operating characteristic”
- AUC = “Area under curve”
- Measure of how well the classes are separated
- Use for “tunable” classifiers (with cut-offs, like Logistic Regression)
- Gives single number $\leq$ 1.0
- Less than 0.5 means “actively bad”

? What does it mean if you have an AUC < 0.5?

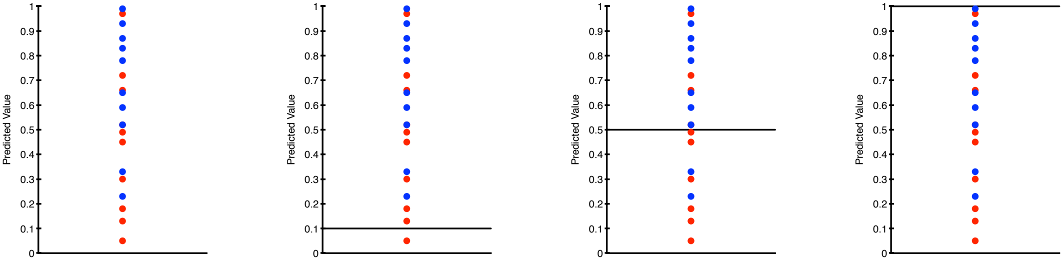
? Pros and cons?

AUC ROC



- Pros
  - Single number
  - Well known
  - Immune to classification threshold
- Cons
  - Only works on binary classification

AUC ROC



- Red = Negative cases
- Blue = Positive cases
- Start at the bottom (or top)
- Sweep up (down)
- Compute the TPR and FPR at each step
- Plot on ROC
- Compute AUC

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 $\langle \varepsilon_1, \varepsilon_2, ..., \varepsilon_n \rangle$, l_p norm defined as:

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

- Common loss functions correspond to various norms:
  - l_1 corresponds to mean absolute error
  - l_2 to mean squared error/least squares
    - Most commonly used
    - Convex
    - Easy to compute
  - 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...

Web Page Link Feature Selection

- Web page link
  - Location
  - Font
  - Size
  - Color
  - ...
- vs. Deep learning
  - Use the raw data
  - E.g. Take a screen shot of the web page
  - Skip the feature engineering phase

Feature Selection Email

I NEED YOUR ASSISTANCE PLEASE Spam ×

Mr Khim Leang khl100@foxmail.com via [redacted]
to [redacted]

Mon, Oct 15, 5:02 PM (1 day ago)

Why is this message in spam? It is similar to messages that were identified as spam in the past.

Report not spam

Dear Friend ,

I am Mr Khim Leang and a personal Accountant/Executive board of Directors with Foreign Trade Bank of Cambodia (FTB).
it is with good spirit of heart i opened up this great opportunity to you A deceased client of mine that shares almost the same name as yours died as a result of heart-related condition on march 2005.His heart condition was duo to the death of the members of his family in the tsunami disaster on the 26 December 2004 in Sumatra Indonesia where they all lost their lives..{More info: http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake_and_tsunami}

There is a draft account opened in my bank in 1999 by a long-time client our bank,a national of your country.he was a CEO/a textile company owner,business man,a miner at kruger mining company here in Cambodia. he was a geologist and consultant to several other mining conglomerates operating in Cambodia,China,Taiwan,Japan,Indonesia,Pakistan,Vietnam all in Asia,before he passed away on 12th march 2005 leaving nobody as the next of kin of his account after his death.

The amount in this account is currently $32,640,000 (Thirty Two Million Six Hundred and Forty Thousand United States Dollars) I want to present you as a beneficiary,I will use my position and influence in our bank to make they release this money to you for us to share.If i wait for days and i do not hear from you,I shall look for another person.

Kindly get back to me for more details

Yours sincerely

...

Mr Khim Leang
Board member
Foreign Trade Bank of Cambodia
Phnom Penh

"Bag of Words"

- Might build a dictionary
  - That is, map from each of m unique words in corpus
  - To a number from $\{1...m\}$
  - Then, each email is a vector $\langle 1, 0, 2, 1, 0, 0, ...\rangle$
  - jth entry is num occurrences of word j
  - Latent Dirichlet Allocation (LDA) uses this approach
  - ? Are there issues with this approach?

“Bag of Words Issues”

- Lose sequence information
  - Use N-grams
  - # of combinations explodes
  - The data dimensionality can get to billions
  - But the feature vector is typically sparse
- Word importance is lost

TF-IDF

- "Term Frequency" – frequency of each word in the document
  - Defined as:

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

- "Inverse Document Frequency" – rareness of the word in the corpus
  - Defined as:

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

- 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?

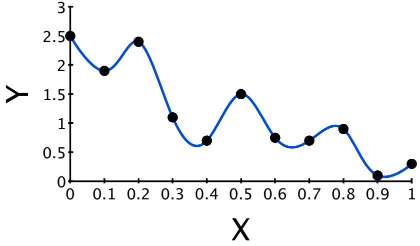
- Country of sender
- Number of words in email
- Time of day sent
- Was the email sent previously? Does it include an email I sent?
- Recipient list disclosed? If not, indicative of spam

Supervised Learning Methodology

- Important to divide available data into
  - Training–used to learn model
  - Validation–used to see if model useful
  - Repeat these two, experimenting
  - Testing–used to evaluate useful models
- Don't touch testing until ready to eval
  - Evaluation on testing must be very last step!
  - ? Why?

Why Test Only Once?

- Overfitting
  - It's easy to build a model that exactly fits your data
  - But doesn't work on new data
  - Maybe the world has changed
    - Can't do much about this
  - Overfitting
    - You don't want to engineer something that only works on your test data



How To Perform Testing

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

How To Perform Testing

- One-off
  - Apply validated model(s), get results
  - Problems?
  - You’re done
  - You’ve used up your test set

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;
```

- Benefits
  - Works for small test sets
  - k can be large

Other techniques for Generating Test Sets

- Jackknife resampling (Leave out one)
- Bootstrap – random sampling with replacement

Questions?