Lecture 18
Wrap-up

Hamed Hasheminia

Agenda

- What model should I use?
- Summary of all the models we learned in this class
- Few models that you can easily learn on your own

What model should I use?

- Ask yourself the following questions:
  - 1. Do I have an output or not?
    - If yes, you need to use a supervised learning technique, otherwise, you must use an unsupervised technique.
  - In this class we mainly focused on supervised learning techniques. (An exception was Principle Component Method)
  - 2. (Assuming you have a supervised learning problem) ask yourself if your output is a quantitative variable or qualitative. If it is quantitative you MUST use one of the regression methods, otherwise you MUST use a classification algorithm.

Here is the list of all Regression models we learned in this class

- Linear Regression Lines
- Lasso & Ridge (on regression lines)
- KNN – Regression
- Decision Tree – Regression
- Bagging, random forest
- Boosting
- ARIMA, ARMA, MA, AR (These can only be applied on time series data)

Here is the list of all Classification models we learned in this class

- Logistic Regression
- KNN – Classification
- Decision Trees (Classification)
- Bagging, Random Forest
- Boosting
- Naïve Bayes
  - GuassianNB for quantitative inputs
  - BernoulliNB and MultinomialNB for categorical inputs

How do we decide which model(s) to use?

- 3. ask yourself if your goal is interpretation or prediction
  - Interpretative models for regression models are:
    - Linear Regression Models Specifically Lasso
    - Simple regression decision trees
  - Predictive regression models are:
    - Random Forest and Boosting
    - KNN (specifically if our feature space is less than 5)
    - ARIMA, ARMA, MA, AR (These can only be applied on time series data). Low order of ARIMA, ARMA, MA, and AR models are relatively interpretable. Higher order are not that interpretable. This is mainly why they are usually only used for prediction.

How do we decide which model(s) to use?

- Interpretative models for classification problems are:
  - Logistic regression (for dichotomous outputs)
  - Classification Decision Trees
- Predictive classification models are:
  - Random Forest
  - Boosting
  - KNN (Specifically if our feature space is less than 5)
  - Naïve Bayes (this is very strong if we have text data)

Advantages and Limitations of Each of the models we learned about

- Linear Regression Models. (advantages)
  - Very interpretable
  - Easy to compute
  - You can run it on Sparse data
  - No need to standardize your data
- Disadvantages
  - Since they assume linear association among variables, it is not that predictable
  - It assumes normally distributed error terms
  - Outliers can easily affect coefficients

Lasso and Ridge Regression

- Ridge Regression
  - Easy to compute
  - Coefficients usually only converge to zero but do not become zero
  - We shall standardize our data
- Lasso Regression
  - Harder to compute – since absolute values do not have first order derivative at 1 point
  - Irrelevant coefficients swiftly become zero
  - We shall standardize our data

Regression KNN algorithms

- Advantages
  - Very easy to compute
  - Intuitive
  - Easily Capture non-linearity
- Disadvantages
  - Results are not interpretable
  - If you have sparse data and feature space is more than 4, this algorithm cannot be used.
  - We shall standardize our data

Simple Decision Trees(regression and Classification)

- Advantages :
  - Extremely interpretable
  - Very easy to train
  - No need to scale your data
- Cons:
  - Not that predictable

Decision Trees Random Forest and Bagging

- Advantages
  - Relatively easy to train
  - Good predictors
  - No need to scale your data
- Disadvantages
  - Not suitable for interpretation

Decision Trees - Boosting

- Advantages
  - Very good predictors
  - No need to scale data
- Disadvantages
  - 3 parameters needed to be Tuned – very time-consuming
  - Easy to overfit

Time series models ARIMA, ARMA, MA, AR

- Advantages
  - AR – good for smoothing patterns
  - MA – good for tackling shocks in the system
  - ARMA – Combines AR and MA – relatively time consuming to train
  - ARIMA – Combines AR and MA, also takes care of linear trends in the model
- Disadvantages
  - Only applicable to time-series data
  - Not that interpretable (Exception is when you have small p and q)
  - You shall clean and prepare data before applying these models.

Naïve Bayes

- Advantages
  - Extremely Fast
  - Great for Text Mining
- Disadvantages
  - Strong Assumption of independency of inputs
  - Terrible for interpretation
  - Never rely on probability predictions

Can we combine few models?

- Yes! We are going to combine few models in our today's Lab session!
- In python, for classification problems we can use:
  - VotingClassifier()
- Can we also combine quantitative methods?
  - The answer is yes again. Generalized Additive Models – GAMs – are developed to deal with this.

What else did we learn in this class?

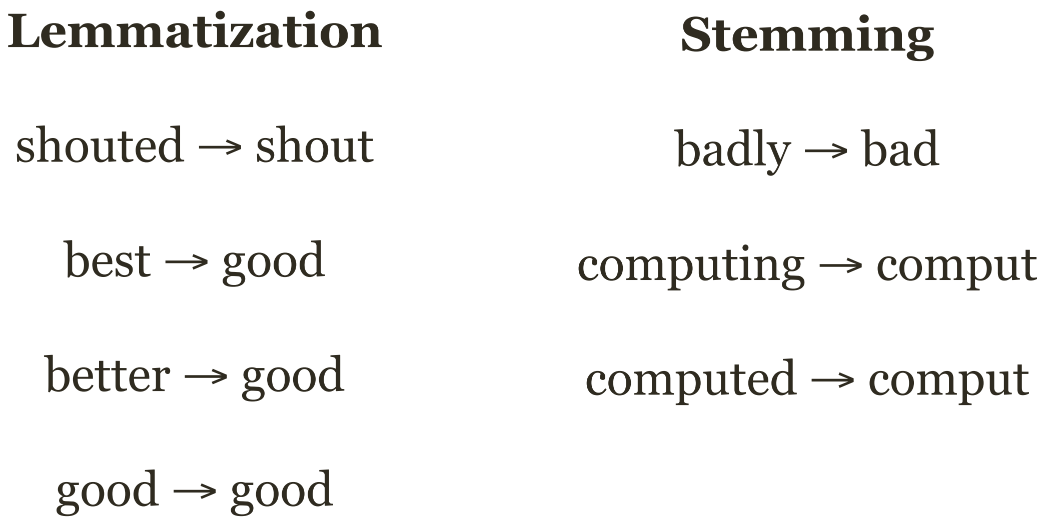
- We discussed how do deal with missing values.
  - We learned wrong ways of dealing with it and correct ways.
  - Here are few wrong approaches:
    - Dropping all missing values if missing values are not random.
    - Using Mean Imputation
    - Using Median Imputation
    - Using Regression Lines
  - Here is a slightly better way of dealing with missing values
    - Regression Lines with error term – this was the best we could do in Python
  - Correct way of dealing with missing values
    - Multiple Imputation

Validation vs Cross-Validation

- Validation: Divide your data to test and train. Train with training Data and Test it with test data. If you have a large dataset, Validation technique is always preferred.
- Cross-Validation: Divides data into few groups. Then train based on all groups but one, and then test it on the one that is left outside. It repeats it on all groups. This way, you are not wasting any Data. Use it when you have a small dataset.
- Can we combine Validation and Cross-Validation?
  - Yes, of course! You can tune based on Cross-Validation and then test it on completely unseen data – test set.

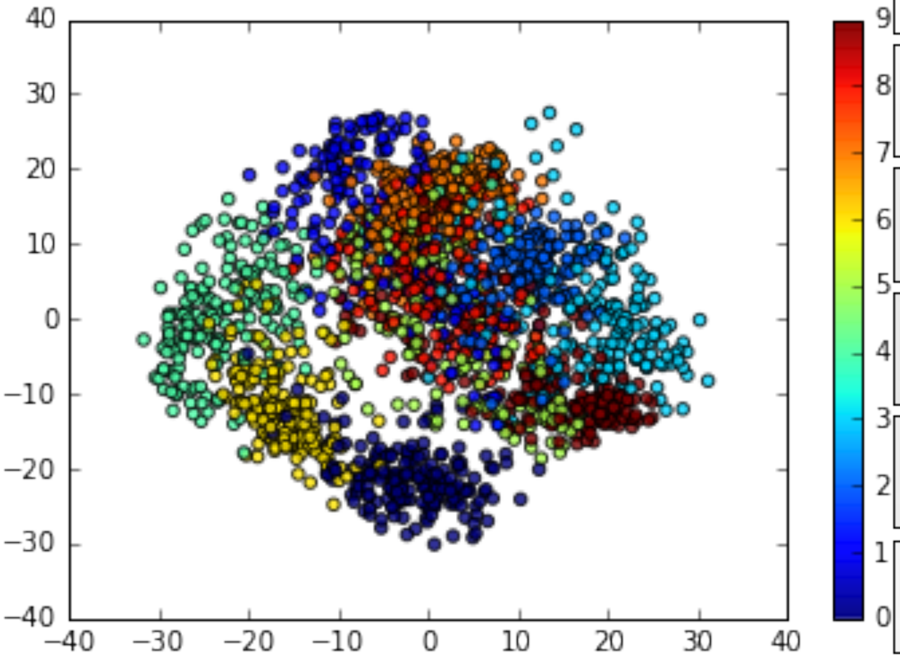
Natural Language Processing

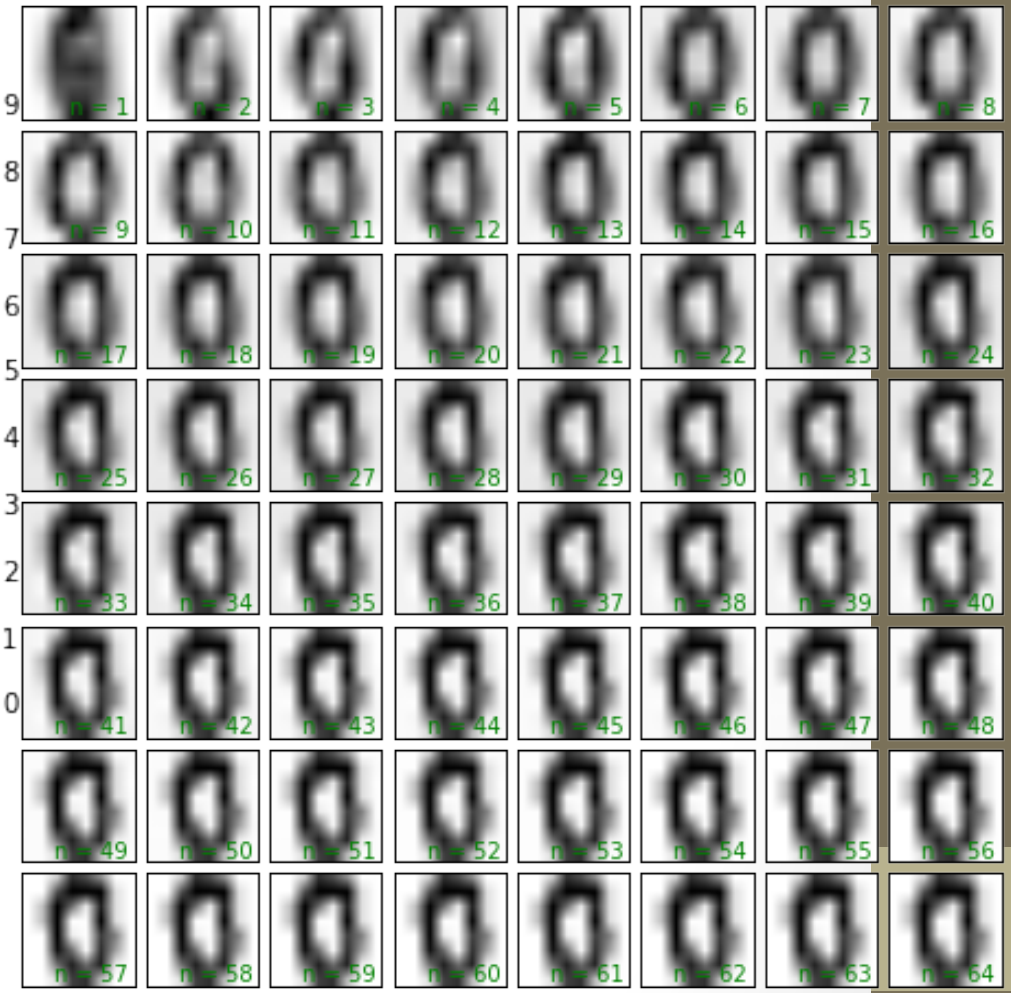
- We learned about stop words
- Bag-of-words (Great of smaller texts)
- TFDIF (Great for long texts)
- Lemmatization and stemming
- We used CountVectorizer() and TfidfVectorizer() in lab. We also used pipeline library.



Lemmatization	**Stemming**
shouted → shout	badly → bad
best → good	computing → comput
better → good	computed → comput
good → good	

Principal Component Method





What is left?

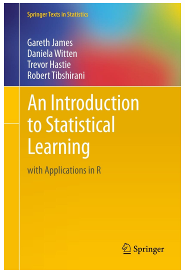
- A lot! But there are few things that I am confident you can learn on your own!
- For Unsupervised Learning Techniques: Easy, Medium, Hard
  - K-Means Clustering (an unsupervised technique to cluster your data.
  - Hierarchical Trees (Another unsupervised technique to cluster data)
  - Principal Curves and Surfaces
  - Kernel Principal Components
  - Self-Organizing Maps (NN)

Few Supervised Learning Techniques left

- For Supervised Learning Techniques: Easy, Medium, Hard
  - Back propagation/feed forward neural Networks
  - Splines
  - Support Vector Machines
  - Ordinal Logistic Regression Models
  - Generalized Additive Models
  - Local Regression Models
  - Directed Graphical Models (Bayes nets)
  - Undirected Graphical Models
  - Linear and Quadratic Discriminant Analysis (LDA, QDA)
  - Fuzzy rule-based systems
  - Markov and Hidden Markov Models
  - Deep Learning (Deep directed networks, Deep Boltzmann Machines, Deep Neural Networks)

If you are interested to learn more, read these three books Sequentially

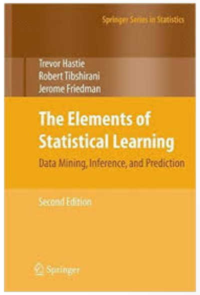
- 1. Statistical Learning with applications in R http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf



- 2. The Elements of Statistical Learning – Data Mining

Inference, and Prediction

https://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

- 3. Machine Learning – A probabilistic Perspective

(By Kevin Murphy)