Tutorial 3

Overview

Non-Parametrics Models

- Definitions
- KNN

Ensemble Methods

- Definitions, Examples
- Random Forests

Clustering

- Definitions, Examples
- k-Means Clustering

Definitions

Definitions

- Parametric Models

Definitions

- Parametric Models
  - Fixed number of parameters - learned (estimated) from data
  - More data ⇒ More accurate models.

Definitions

- Parametric Models
  - Fixed number of parameters - learned (estimated) from data
  - More data $\Rightarrow$ More accurate models.
- Non-parametric Models

Definitions

- Parametric Models
  - Fixed number of parameters - learned (estimated) from data
  - More data ⇒ More accurate models.
- Non-parametric Models
  - Number of parameters grows with the amount of data
  - More data ⇒ More complex models.

Definitions

- Parametric Models
  - Fixed number of parameters - learned (estimated) from data
  - More data ⇒ More accurate models.
- Non-parametric Models
  - Number of parameters grows with the amount of data
  - More data ⇒ More complex models.
- Parametric or Non-parametric? What are the parameters?
  - Decision Trees
  - Naive Bayes
  - KNN
  - Random Forests
  - K-Means Clustering

k-Nearest Neighbour

k-Nearest Neighbour

- How does it work?

k-Nearest Neighbour

- How does it work?
- What is the effect of k with respect to the fundamental tradeoff in machine learning?

k-Nearest Neighbour

- How does it work?
- What is the effect of k with respect to the fundamental tradeoff in machine learning?
- What is the runtime of a naive implementation? How could you speed this up?

Ensemble Methods

Ensemble Methods

- Learning algorithms that take classifiers as input and use the output of each classifier to determine a classification

Ensemble Methods

- Learning algorithms that take classifiers as input and use the output of each classifier to determine a classification
- Averaging
  - Take the average of the outputs of each classifier (or mode if categorical)

Ensemble Methods

- Learning algorithms that take classifiers as input and use the output of each classifier to determine a classification
- Averaging
  - Take the average of the outputs of each classifier (or mode if categorical)
- Bagging
  - Each classifier in the ensemble votes on an output with equal weight
  - Each classifier is trained with a random subset of the training set

Ensemble Methods

- Learning algorithms that take classifiers as input and use the output of each classifier to determine a classification
- Averaging
  - Take the average of the outputs of each classifier (or mode if categorical)
- Bagging
  - Each classifier in the ensemble votes on an output with equal weight
  - Each classifier is trained with a random subset of the training set
- Boosting
  - Incrementally build the ensemble. When training new models higher weight is given to data that was mis-classified by previous models

Ensemble Methods

- Learning algorithms that take classifiers as input and use the output of each classifier to determine a classification
- Averaging
  - Take the average of the outputs of each classifier (or mode if categorical)
- Bagging
  - Each classifier in the ensemble votes on an output with equal weight
  - Each classifier is trained with a random subset of the training set
- Boosting
  - Incrementally build the ensemble. When training new models higher weight is given to data that was mis-classified by previous models
- Stacking
  - Train a classifier to combine the predictions of the other classifiers

Ensemble Methods

- Learning algorithms that take classifiers as input and use the output of each classifier to determine a classification
- Averaging
  - Take the average of the outputs of each classifier (or mode if categorical)
- Bagging
  - Each classifier in the ensemble votes on an output with equal weight
  - Each classifier is trained with a random subset of the training set
- Boosting
  - Incrementally build the ensemble. When training new models higher weight is given to data that was mis-classified by previous models
- Stacking
  - Train a classifier to combine the predictions of the other classifiers
- And more!

Random Forests

Random Forests

- How do they work? How do you train them?

Random Forests

- How do they work? How do you train them?
  1. Create several bootstrap samples of the data

Random Forests

- How do they work? How do you train them?
    1. Create several bootstrap samples of the data
    2. Train a **random** decision tree on each bootstrap sample

Random Forests

- How do they work? How do you train them?
  1. Create several bootstrap samples of the data
  2. Train a **random** decision tree on each bootstrap sample
  3. Test by averaging the predictions of each tree

Random Forests

- How do they work? How do you train them?
  1. Create several bootstrap samples of the data
  2. Train a **random** decision tree on each bootstrap sample
  3. Test by averaging the predictions of each tree
- How does the number of trees affect the fundmental tradeoff of machine learning?

Random Forests

- How do they work? How do you train them?
  1. Create several bootstrap samples of the data
  2. Train a **random** decision tree on each bootstrap sample
  3. Test by averaging the predictions of each tree
- How does the number of trees affect the fundmental tradeoff of machine learning?
- How does the amount of randomness in the trees affect the fundamental tradeoff of machine learning?

Clustering

Clustering

- An unsupervised method - not given labels, but want to learn something about the data
  - Specifically the classes, or groups, that the data falls into

Clustering

- An unsupervised method - not given labels, but want to learn something about the data
  - Specifically the classes, or groups, that the data falls into
- Classes are determined by similarty between data and dissimilarity to other classes

Clustering

- An unsupervised method - not given labels, but want to learn something about the data
  - Specifically the classes, or groups, that the data falls into
- Classes are determined by similarty between data and dissimilarity to other classes
- e.g. Types of genes, variants of a disease, topics on Wikipedia, friends on Facebook, etc.

k-Means

k-Means

- How does it work?

k-Means

- How does it work?
- K++ means - what problem does this address?

k-Means

- How does it work?
- K++ means - what problem does this address?
- Label switching problem