Feature Selection and Extraction

UW ECE 657A - Core Topic

Mark Crowley

Lecture Outline

Feature Extraction vs. Feature Selection

- Given a dataset $X : N \times D$ matrix.
- We can **extract** or **transform** new features F to describe the variation, distances, proximity, etc, in the data such that $|F| < D$.
- We can **select** from the existing features the ones which are the most representative F to describe X for $|F| < D$.

Dimensionality Reduction

Can be seen as a preprocessing step that may be required for:

- Reducing costs of training and learning algorithms
- Reducing storage and future measurement costs
- Visualization of the data to *debug* or otherwise *understand* your data.
- Finding the **intrinsic dimensionality** (features)
  - Many applications have a large number of features that may be redundant, irrelevant or sparse (e.g text documents where features are words)
- If the number of samples are much smaller than number of dimensions then this makes learning at a desired resolution difficult $\rightarrow$ **The Curse of Dimensionality**.

Motivation: The Curse of Dimensionality

- A term introduced by Richard Bellman to illustrate the explosion of the samples required when we have higher dimensions.
- The number of samples required to fully represent a 10 valued function with one dimension=10
- For 20 dimensions with same number of values (10 datapoints in each dimension) we will need 10^{20}
- The number of samples **grows exponentially** in terms of the dimensions.
- This poses a problem for **non-parametric** algorithms such as most classification algorithms that require many training samples.

Motivation: The Curse of Dimensionality

Learn a 1-D function from data:

How many data points are needed?

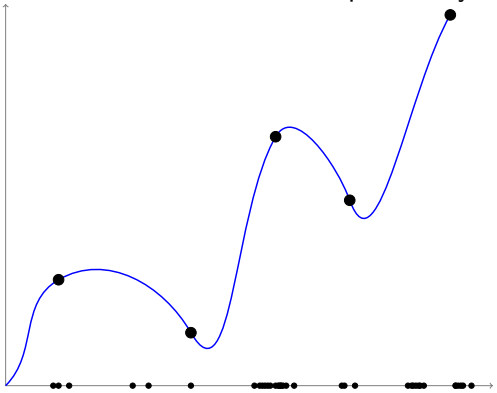
Motivation: The Curse of Dimensionality

Learn a 1-D function from data:

How many data points are needed? N points

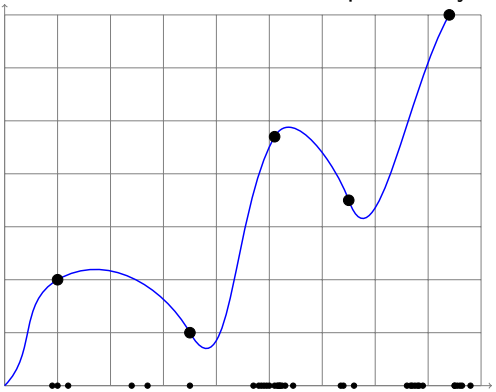
Motivation: The Curse of Dimensionality

What if the data is better explained by a 2-D function?



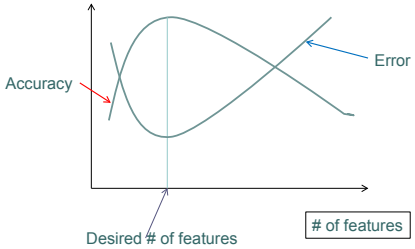
Motivation: The Curse of Dimensionality

What if the data is better explained by a 2-D function? N^2 points



Motivation: Peaking Phenomenon

- For a fixed number of samples increasing the number of features may degrade the performance. (eg. genetics data: very high dimensional, small data sample)
- Classic solution is to assume fixed order of features, only choose most important.
- Newer approach: use classification algorithms to reduce feature space (see https://www.sciencedirect.com/science/article/abs/pii/S0167865508001426



Dimensionality Reduction and Feature Extraction Methods

Feature Extraction and Feature Selection

- Feature Extraction combines original features into a new set of features by transformation or projection.
- Ideally the transformation or the projection is according to an objective that helps finding the significant set of new features of lower dimensions than the original set of features
- Sometimes the transformation produce new features that may have some relevant meaning to the data but many projection cases produce features that are difficult to interpret or relate to the meanings of the original features.

Feature Extraction

- **Feature Extraction:** ways to transform the original features in the data into new features which have some advantage such as
  - lower dimensionality
  - better description of the variance in the data
  - better ability to discriminate (distinguish) data points or clusters of data points

Feature Selection

- **Feature Selection** is attempt to find a subset of the original features which satisfy some criteria.
- Ideally the selected subset includes the significant features and eliminates irrelevant and redundant features.
- The selected features being a subset of the original are interpretable and keep their original meaning.
- Approaches include:
  - Feature Ranking
  - Subset Selection which includes:
  - Wrapper approach and Filter approach
  - Both use search strategies such as:
    - Exhaustive
    - Sequential (Forward or Backward)

Lecture Outline

General Approach

- Search the feature space for a *subset of features* that optimizes a selection criterion, an objective function (ie. a quality index or a classifier output).
- Challenge: If we have n samples and d features then there are 2^d possible subsets of d.
- Components:
  1. Selection Criteria
  2. Search Strategy

Feature Ranking

We need to use more feasible methods.

Idea: Treat the problem as a search problem and find ways to reduce the search space.

Feature Ranking:

1. (also called Best Individual Features (BIF) and Naive Selection)
2. Evaluate the individual features according to a quality measure (how good the feature is for discriminating the class)
3. Sort the features according to their value of the quality measure
4. Select the best m features

Advantage: search complexity is $O(m)$ after sorting. It can be used for very large number of features

Problem: features are considered in isolation

Limitations of Correlation

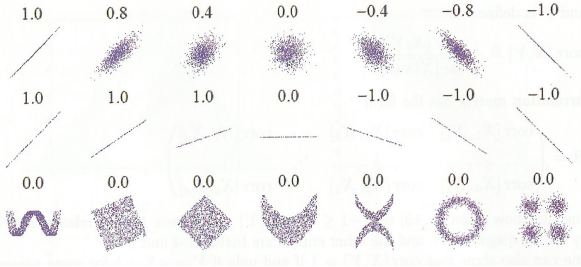


Figure: From (Murphy, 2012) Fig 2.12

Quality Measures for Feature Ranking

Some quality measures include:

- Mutual Information: between a feature and class
- Information Gain
- Chi-squared test (χ^2)

Recall: Mutual Information (MI)

The **mutual information (MI)** between two vectors X, Y measures how similar the joint distribution $p(X, Y)$ is to the factored distribution $p(X)p(Y)$:

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

- MI(X,Y) is always nonnegative
- Equals 0 iff X, Y are independent
- Notice this is just the KL-Divergence between the distributions $p(X, Y)$ and $p(X)p(Y)$

Mutual Information For Feature Ranking

In our case, mutual information is:

$$MI(X_f, Y) = \sum_{x_{fi} \in X_f} \sum_{y_k \in Y} p(x_{fi}, y_k) \log \frac{p(x_{fi}, y_k)}{p(x_{fi}) p(y_k)}$$

- where $x_{fi} \in X_f$ is the value of feature (column) f for datapoint i
- and $y_k \in Y$ are the class labels
- a feature with high MI may not have high probability but it's better at identifying the classes.

Feature Selection Methods

- Feature Ranking
  - Quality Measures for Feature Ranking
- Choosing Feature Subsets
  - Filter Approach: evaluate feature subsets based on information content
  - Wrapper Approach: evaluate feature subsets based on maximizing accuracy for final classifier being used
  - Objective Functions and Distance Measures: Embedded vs. Hybrid
- Search Strategies
  - Exhaustive Search
  - Sequential Forward Selection
  - Sequential Backward Selection
  - Sequential Floating Selection
  - Oscillating Search

Lecture Outline

Feature Extraction Methods

- Feature Extraction combines original features into a new set of features by **transformation** or **projection**.
- Projection is according to some objective function or concept.
  - objective should find the *most significant set of new features*
  - used as a **lower dimensional subspace** than the original set of features
- Interpretation:
  - Transformation can produce new features have interpretable meaning to the data
  - But many projections produce features that are difficult to relate to the original features.

Methods of Feature Extraction

- Linear methods:
  - Unsupervised (no class label information, e.g .PCA, ICA)
  - Supervised (class labels know, e.g LDA)
- Nonlinear methods
  - Global (preserves global properties, eg. MDS, Isomap, Kernel PCA)
  - Local (preserves properties within local neighborhood, e.g. LLE)