**CS286 Solving Big Data Problems – Exam #2 Study Guide**

**By: Zayd Hammoudeh**

**Lecture #14 – Introduction to Data Science**

**Features**

|  |  |  |
| --- | --- | --- |
| **Feature –** Individual, measurable property of an observation. In supervised learning, it is used for predicting the target.  **Additional Names:**   * **Predictor** * **Attribute** * **Variable**   **Vectorizing Data** – Given an observation, assign features to dimensions. | **Attribute Types**   * **Continuous** (Numeric) – **Data across a continuum**. * **Binary** – Two possible values. * **Categorical** – Finite set of possible values. * **Text** –Vocabulary or n-gram.   + **Bi-gram** – 2 consecutive words. | **Qualities of Good Features/Feature Sets**   * **Correlation** with the target (i.e. class value) * Features are **independent** * **Number of Features in Minimized** * **Number of features is much less than the number of data points**. * **Features do not leak the target** meaning they do not provide information that would not be available under normal circumstances. |

|  |  |  |
| --- | --- | --- |
| **Set Dimensionality**   * – Number of observations * – Number of Features   Ideally: | **Principal Component Analysis Procedure** – **Dimensionality Reduction** | |
| **Procedure**   1. **Calculate covariance matrix of the original input matrix.** 2. **Perform eigenvalue decomposition of the covariance matrix.** 3. **Look for orthogonal dimensions with the greatest variance (principal)** 4. **Project input matrix into those dimensions.**   **Relative covariance preserved across projection**. | – Representation of the total variance of the principal components.  **Number of principal components and eigenvectors equivalent to the number of dimensions in the original input matrix.**  **Since the data is mapped to new dimensions, not easy to map principal components to the original data.** |

|  |  |  |
| --- | --- | --- |
| **Feature Scale** | | **Target Leaks** – **Sample data has “extra” information/features not found outside the sample**. May cause the model to be unrealistically good.  **Bias**   * **Sampling Bias** – Non-representative sample * **Observer Bias** – Corrupted/invalid measurement of features * **Funding Bias** – “Fudging”/manipulating of the data.   **Simpson’s Paradox** – **Trend that appears in different groups disappears (or changes) groups are combined**. **Example**: The average of averages does not equal the average all points. |
| **Potential Issues**   * **Relative Magnitude** – **Some features may have different magnitude than others** causing those features to dominate. * **Absolute Magnitude** – When features are too large or too small, **data may be lost to overflow or underflow** respectively. | **Potential Solutions**   * **Standardization** –Transform data to measure how many standard deviations it is from the mean. * **Normalization** – Transform data such that the range is between 0 and 1. |

**Algorithms and Evaluation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Applications of Clustering**   * **Image Segmentation** * **Dendrogram (including Phylogentic Dendogram)** * **Market Segmentation** | **K-Means Clustering Algorithm**   1. **Initialize the K centroid** 2. **Assign points to the nearest centroid.** 3. **Update the centroids.** 4. **Check Stop Conditions** | **Methods for Selecting the Initial Clusters**   1. **Random** 2. **El Agha et. al. Method** 3. **All at the same point (e.g. origin)**   **Stop Condition Options**   1. **Min change of median from last iteration**. 2. Sufficiently **small intra-cluster distance** 3. Sufficiently **large intercluster distance** | **Intracluster Distance** – Distance between **points within the same cluster**. **Measures the cluster’s similarity**.  **Intercluster Distance** – **Distance between** **the centroids**. **Measures data separation between clusters.** |

**Up to slide 20 on lecture #14.**