Good features

- Informative
- Independent
- Monotonic with predictive probability

Informative

- Mutual Information measures how informative a feature $X$ is about label $Y$
  - $I[X; Y]$: $X$'s information about $Y$
  - $H[X|Y]$: $X$'s extra randomness not rel. to $Y$
  - $H[Y|X]$: $Y$'s extra randomness not rel. to $X$
  - $H[X] = H[X|Y] + I[X; Y]$

Informative features

- Examples
  - A pixel of an image for predicting object type
  - Income level for predicting debt default
  - Having the word "great" in a review for predicting sentiment of a review

Independent features

- $X_1$ and $X_2$ are independent variables given $Y$
  - $I[(X_1, X_2); Y] = I[X_1; Y] + I[X_2; Y]$

- $X_1$ and $X_2$ are dependent variables
  - $I[(X_1, X_2); Y] < I[X_1; Y] + I[X_2; Y]$
  - Extreme case: $X_1 = X_2$

Monotonic with predictive probability

- Feature monotonically increases (or decreases) with the predictive probability

- Each to apply threshold
  - Simpler structure for trees
  - Tend to have linear decision boundary
  - Distance is meaningful
Problem: too many features

- Classifiers have large chance to use noisy features
- Example:
  - 10 training instance, 5 pos and 5 neg
  - A feature strongly correlated with label, e.g. correlation coeff = 1.0
  - What’s the chance that a random feature appears to be strongly correlated?
    - Answer: 1/1024

Feature normalization

- Feature Normalization/Standardization
  - Centering: subtracting feature mean from each instance
  - Normalization: dividing each instance by feature std-s

Categorical features

- Convert categorical features to continuous ones
  - Not losing information
  - Smooth in terms

Feature Pruning

- Feature Pruning (Selection)
  - Remove redundant features
  - Remove features with little information about the label
- Methods
  - Remove features with low mutual information
  - Searching feature combinations by classifier performances

A case study

- Extract a feature vector from a document (a list of words)

A case study

- Bag of words

http://mlg.postech.ac.kr/research/nmf
A case study

- **Sum of Word Embeddings**
  - Word embeddings (not in this course): learning a vector representation $v_j$ for each word $j$ in the vocabulary

  ![Word Vectors](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/06/06062705/Word-Vectors.png)

- Sum of Word Embeddings
  - Sum up word embeddings in a document $d = \{v_1, \ldots, v_k\}$

  $\text{FeatOf}(d) = \sum_{k=1}^{K} v_{j_k}$