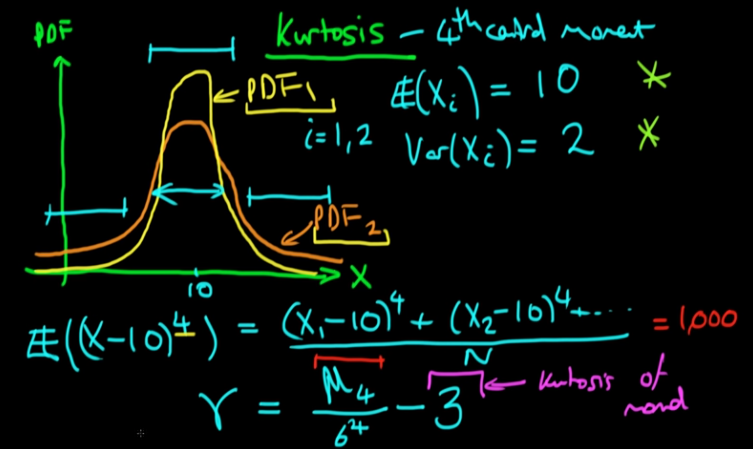
Feature Transformation

* Definition: The problem of pre-processing a set of features to create a new (smaller? More compact?) feature set, while retaining as much (relevant? Useful?) information as possible
  + Consistent with feature selection, but selection is a subset of feature transformation
  + m < n (almost always)
  + Linear feature transformation: P^T\*x. linear combinations of original features projecting original features into new feature space
    - X1, x2, x3, x4
      * Feature selection: x1, x2
      * Feature transformation: 2x1+x2
  + Implicit assumption is that we don’t need all n features
* Why do feature transformation?
  + We’ve kind of been doing it all along (SVM, NN), why do it separate?
  + Ex: information retrieval (ad hoc retrieval – google problem)
    - Google problem – database somewhere with a bunch of documents, want to retrieve subset of documents with some relevant query
    - Hard because you don’t know what the query up front will be
  + What are our features?
    - Words
      * Problems
        + Lot of words – curse of dimensionality
        + Words mean multiple things – polysemy

False positives

* + - * + Can say the same thing with many words – synonomy

False negatives

* + - * + How to combine words together to solve \*my like problems?
* Principal Components Analysis
  + (Particular example of something called an eigenproblem)
  + Eigenvectors and eigenvalues
  + Properties
    - Finds directions of maximal variance
    - Finds directions that are mutually orthogonal (constraint)
      * Means it is a global algorithm
    - Gives you the best reconstruction
      * Can reconstruct all original data given new dimensions, no throwing away information. Minimizes L2 error (throwing away one of the two new dimensions – axis 2 in lecture).
    - Eigenproblem
      * Associated with each one of the new dimensions is its eigenvalues. You can throw away the ones with the least eigenvalue, which is analogous to throwing away the features with the least variance. 0 is ignorable, will not affect reconstruction error. 0 variance is 0 entropy, which means it is irrelevant, but maybe useful
    - In practice, people subtract the mean of the data which will center the data around the origin. This means that the new dimensions will now go through the origin.
    - Well studied – fast.
    - Not clear that throwing away data would help with classification
    - Like filtering
    - Takes the data, finds different set of axes and lines up variance of data with new axes so that we can drop the least significant ones which gives us a way to do feature selection, but the whole thing is a feature transformation algorithm in that it first moved the data around to be able to do that.
      * You do transformation into a new space where you know how to do filtering
  + PCA is about finding correlation by maximizing variance which allows you reconstruction
  + PCA does a terrible job at solving the blind source separation problem
  + PCA is not directional
  + PCA will find global features
* Independent Components Analysis
  + Tries to maximize independence
  + Tries to find linear transformation for feature space into new feature space such that each of new features are mutually independent (in a statistical sense - their mutual information is equal to zero)
    - [x1, x2, …, xi] => [y1, y2, …, yi]
    - I(y\_i, y\_j) = 0
    - I(y;x) = great as possible
  + Fundamental assumption: given observables, try to find hidden variables under the assumption the hidden variables are independent of each other
    - Cocktail party problem – in a large group of people and trying to listen to a conversation, how do you pull out one source that you care about while separating out the ones you don’t care about
      * Source are arbitrarily linearly combined. Can use ICA to separate them.
  + How to do this with actual numbers
    - Matrix - Rows represent features, columns represent samples
    - Mutual information – how much one variable tells you about another
  + Does a great job at solving the blind source separation problem.
  + ICA is highly directional
  + ICA will find local features (finds edges in images).
* PCA vs ICA
  + PCA – mutually orthogonal (this makes it a global algorithm)
  + ICA – mutually independent (although PCA finds independent things by maximizing variance – finds a bunch of orthogonal gaussians).
  + PCA – maximal variance
  + ICA – Maximize mutual information between all the original features together and the new features set
  + PCA – ordered features
  + PCA, ICA – bag of features
* Central limit theorem
  + Allows us to solve problems and make inferences using the normal distribution even when the population is not normally distributed
  + Common rule of thumb for sample size is at least 25 samples?
* Kurtosis
  + 4th central moment of a distribution
  + Gives more insight into distribution past the mean and variance  
    
  + Subtract 3 because this is the normal distribution
  + Tells us how much “fatter” or thinner” than the normal distribution it is. Positive gamma = fatter and vice versa
* Alternatives
  + RCA – random components analysis
    - Will pick any direction and project the data out in that direction.
    - Works remarkably well if the next thing you are going to do is some kind of classification
    - N->m, m<<n. Works well since it picks up some correlations on much less data
    - Will probably project down to a lower dimensional space, but probably not as far as PCA or ICA
    - You could project into higher dimensional spaces as well
    - The big advantage is speed (it is also cheap, simple and easy).
      * “you must earn your complexity”
  + LDA – linear discriminate analysis
    - Finds a projection that discriminates based on the label (feels a lot like supervised learning).
    - Finding linear separators between clumps of points
    - Pays attention to how the features will be used (with regards to the label)
* Summary
  + A is for analysis
    - PCA, ICA, RCA, LDA
  + Relations between different transformation algorithms
  + Analysis of the data
    - ICA will tell underlying structure of the data, more probabilistic and coincidentally about linear algebra
    - PCA more of a linear algebraic model and sometimes coincidentally about probability
    - Linear algebra is easier to do and think about and cheaper to execute, less prone to local minima but often not quite the answer you want
    - Probabilistic method is harder but often does give the answer you want