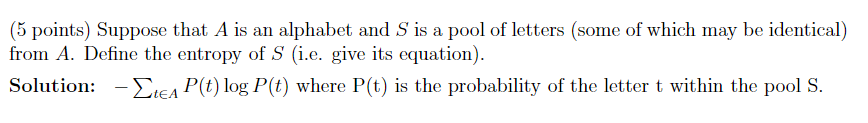
Midterm 2 possible questions



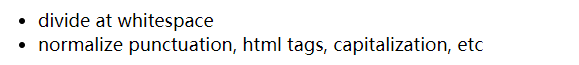
* Word types vs. word tokens

word type: a dictionary entry such as "cat"

word token: a word in a specific position in the text

* **For n words, there are n2 possible bigrams. But our amount of training data stays the same. So the training data is even less adequate for bigrams than it was for single words.**
* Tokenization

Our first job is to produce a clean string of words, suitable for use in the classifier.



**Format features such as punctuation and capitalization may be helpful or harmful, depending on the task.**

So "normalize" may either involve throwing away the information or converting it into a format that's easier to process.

* stemming (Porter stemmer)

Stemming converts inflected forms of a word to a single standardized form. It removes prefixes and suffixes, leaving only the content part of the word (its "stem")

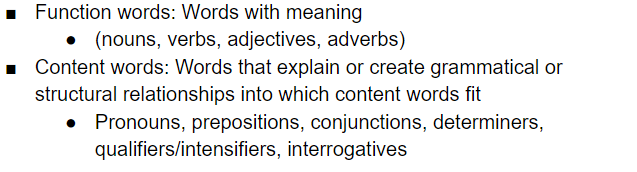
i.e. help, helping, helpful -> help

* stop words & rare words

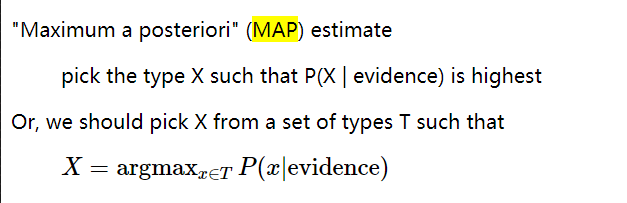
Very frequent words ("stop words") often convey very little information about the topic. So they are typically deleted before doing a bag of words analysis.

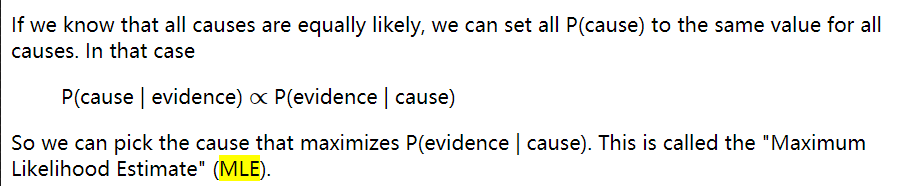
Rare words may be deleted. More often, all rare words are mapped into a single placeholder value (e.g. **UNK**). This allows us to treat them all as a single item, with a reasonable probability of being observed.

* Backchannel: occurs when one participant is speaking and another participant interjects responses to the speaker (“hmm”, “uh-huh”, “yeah”)
* function vs. content



* MAP and MLE versions of the estimation equations

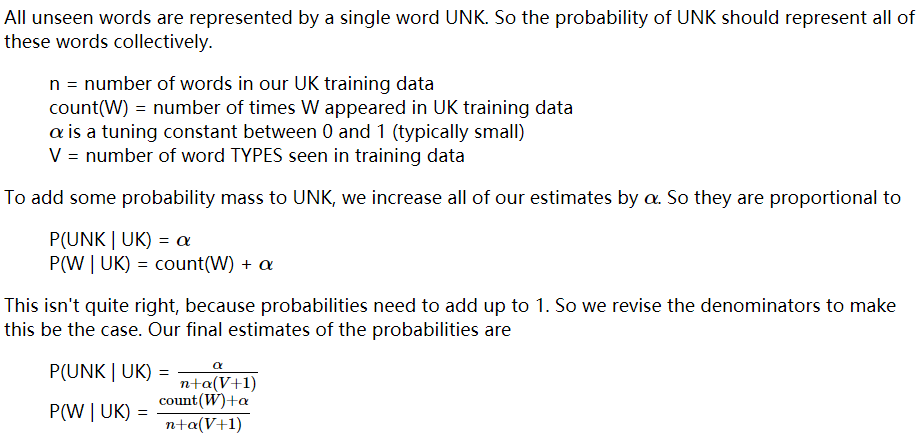




**The MLE estimate can be very inaccurate if the probabilities of different causes are actually different and not the same value as was assumed. On the other hand, it can be a sensible choice if we have no information about the prior probabilities or if there is reason the believe P(Cause) should be constant across all possible causes.**

* **Avoiding overfitting ---smoothing!!!**

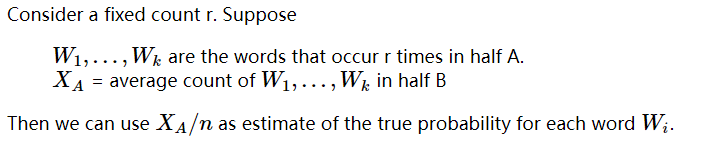
for uncommon words: Words that didn't appear in the training data get estimated zero probability. Words that were uncommon in the training data get inaccurate estimates.



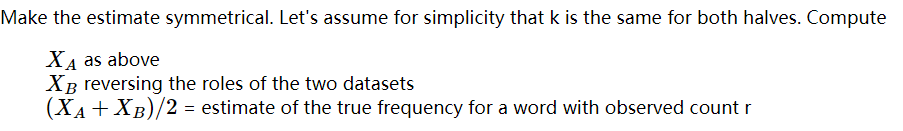
**overestimates probability of unseen words**

**underestimates probability of common words**

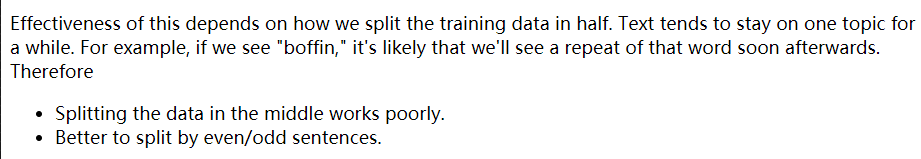
* Deleted estimation: **Also called cross-validation or held-out estimation.**



**Tweaks on deleted estimation:**



**Text tends to stay on one topic for a while.**



***Performance of Deleted Estimation:***

**Infrequent words overestimate their true probability. Estimates for common words are usually accurate.**

* N-gram smoothing

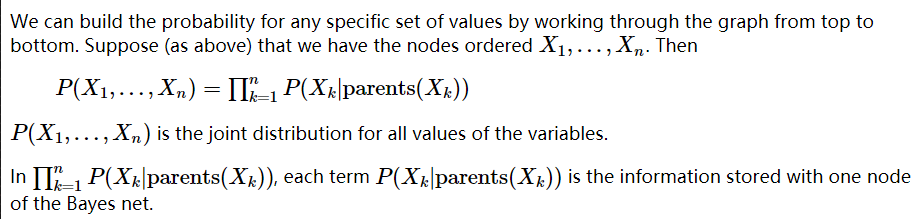
Idea 1: If we haven't seen an ngram, **guess its probability from the probabilites of its prefix (e.g. "the angry") and the last word ("armadillo").**

Idea 2: **Guess that an unseen word is more likely in contexts where we've seen many different words. I.e. some contexts are much more predictable (“hand me the …” can be followed by anything but “I went to mcdonalds and bought a …” is usually followed by “hamburger”.**

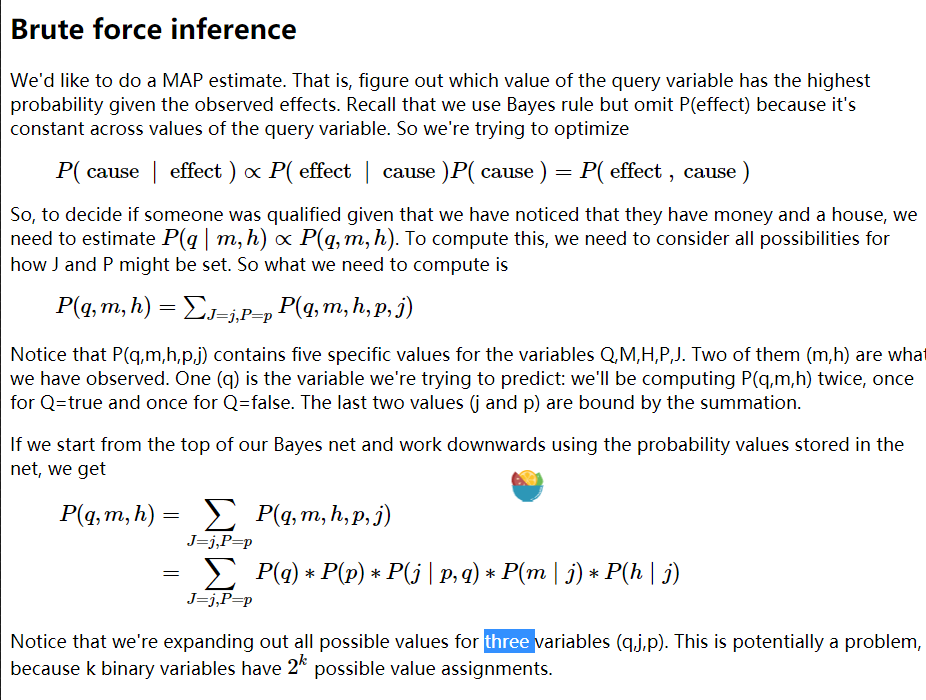
A high-level point: conditional independence isn't some weakened variant of independence. Neither property implies the other.

**A Bayes net is a directed acyclic (no cycles) graph (called a DAG for short). So nodes are partially ordered by the ancestor/descendent relationships.**

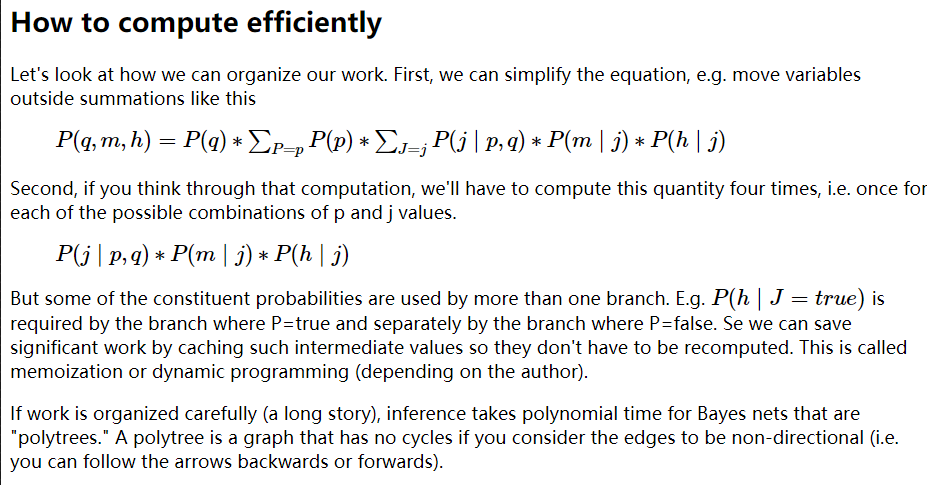
**Each node is conditionally independent of its ancestors, given its parents. That is, the ancestors can ONLY influence node A by working via parents(A)**



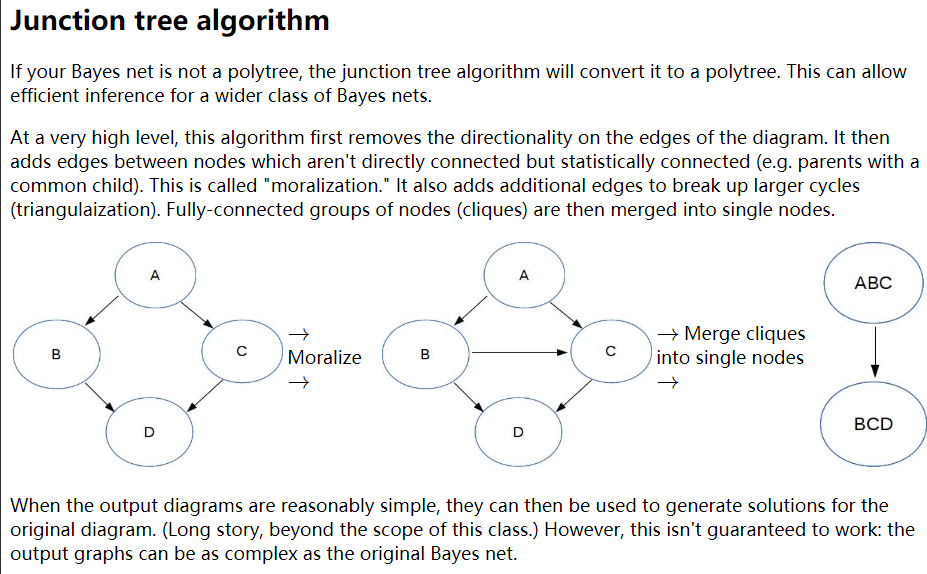
**the choice of variable order affects whether we get a better or worse Bayes net representation. We could imagine a learning algorithm that tries to find the best Bayes net, e.g. try different variable orders and see which is most compact. In practice, we usually depend on domain experts to supply the causal structure.**



**Notice that we're expanding out all possible values for three variables (q,j,p). This is potentially a problem, because k binary variables have 2k possible value assignments.**

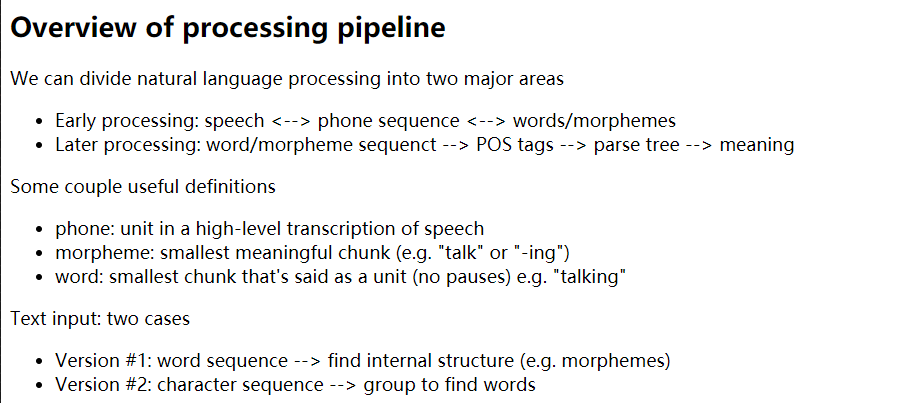


**Bayes nets can be computed in polynomial time when there are polytrees - which means that when you remove the direction from the graph you’re still left with no cycles**

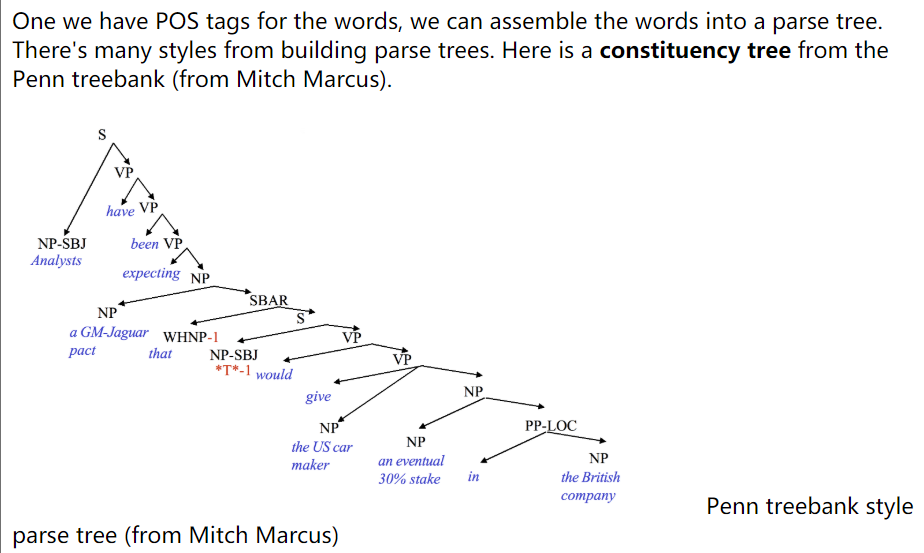


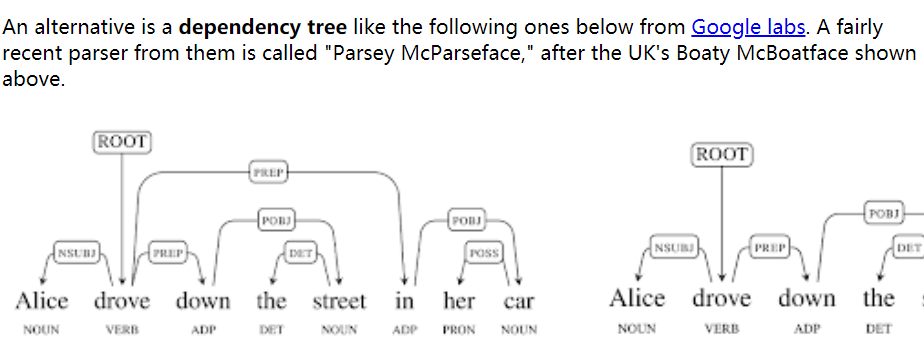
**At a very high level, this algorithm first removes the directionality on the edges of the diagram. It then adds edges between nodes which aren't directly connected but statistically connected (e.g. parents with a common child). This is called "moralization." It also adds additional edges to break up larger cycles (triangulaization). Fully-connected groups of nodes (cliques) are then merged into single nodes.**

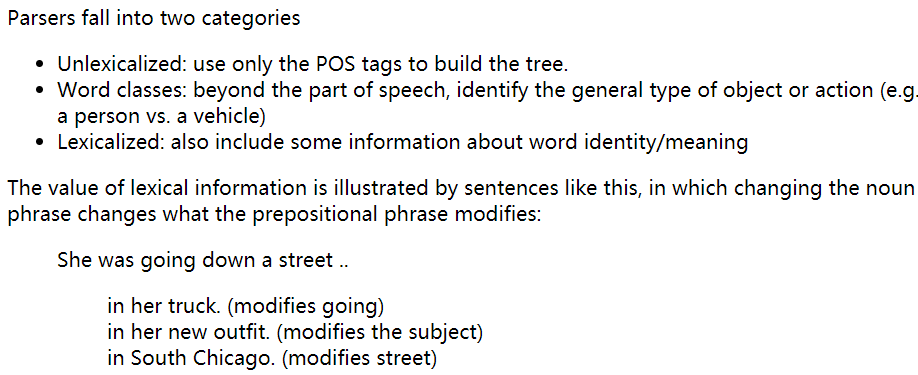
**NLP**



**constituency tree & dependency tree**





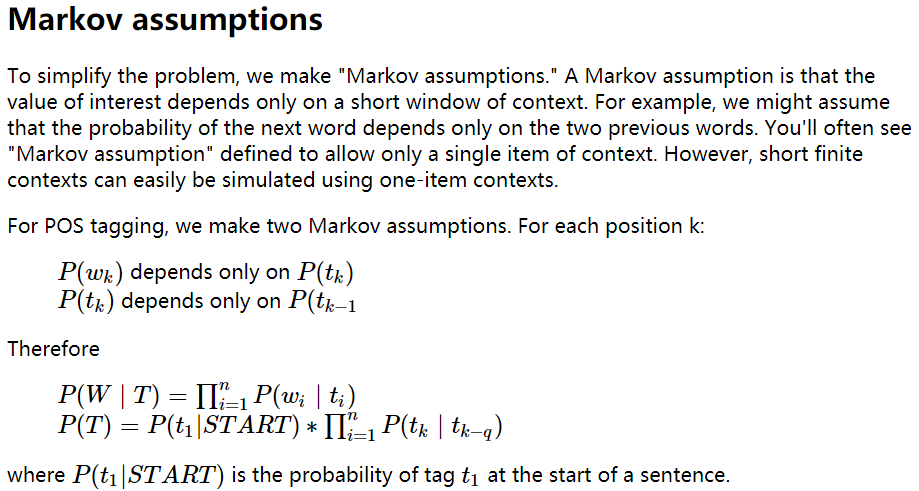


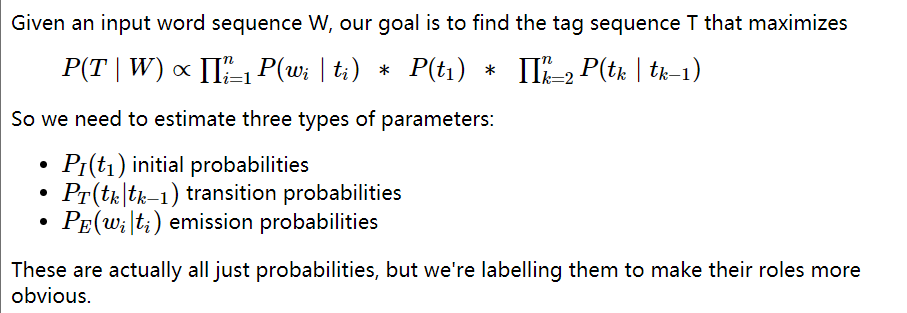
baseline algorithm might pick the most common tag for each word, ignoring context. In the table above, the most common tag for each word is shown in bold: it made one error out of six words.

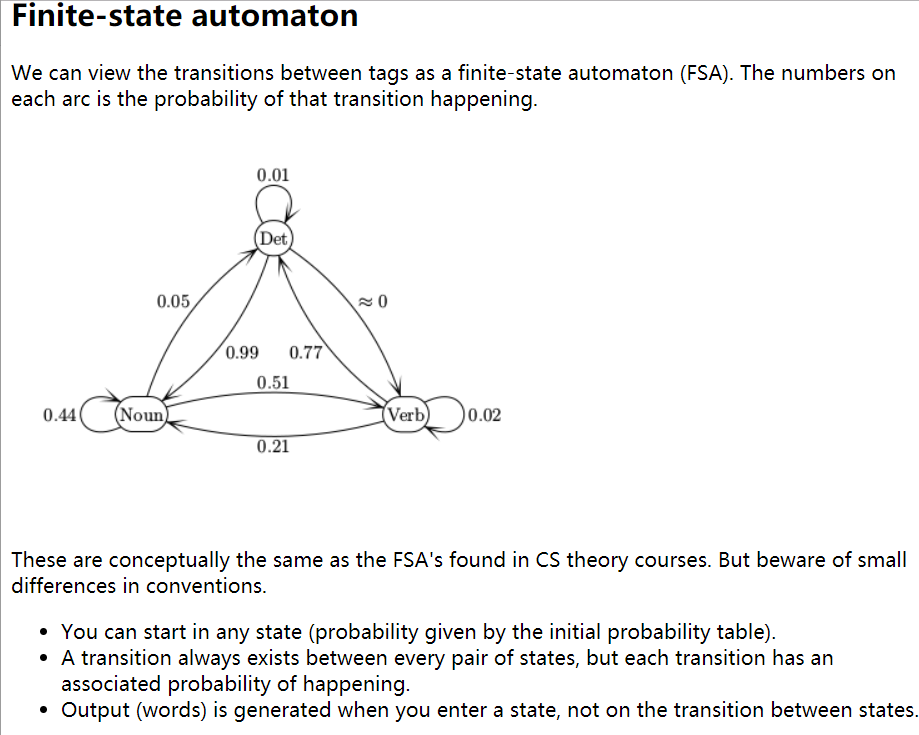
**The test/development data will typically contain some words that weren't seen in the training data. The baseline algorithm guesses that these are nouns, since that's the most common tag for infrequent words.** An algorithm can figure this out by examining new words that appear in the development set, or by examining the set of words that occur only once in the training data ("hapax") words.

Hapax： words only occurs once

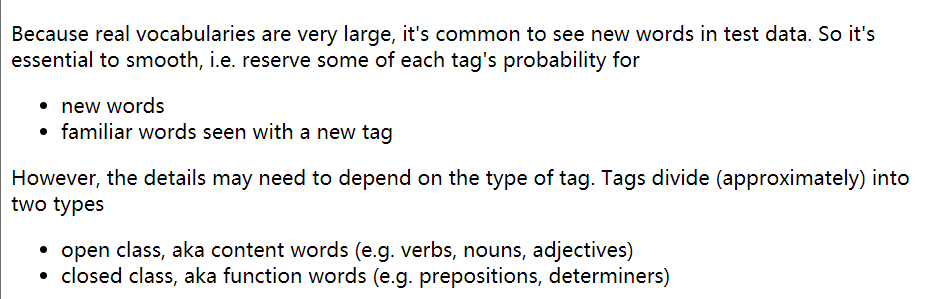
**A Markov assumption is that the value of interest depends only on a short window of context.**



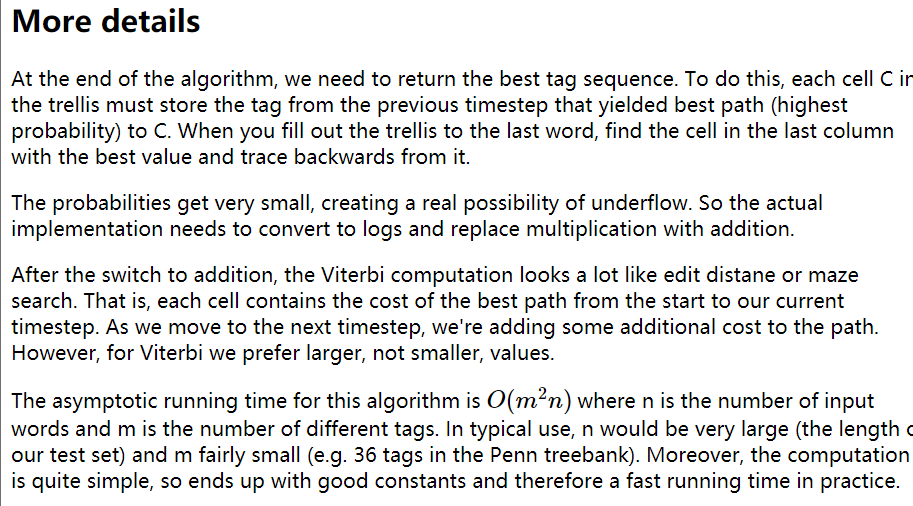




**Output (words) is generated when you enter a state, not on the transition between states.**



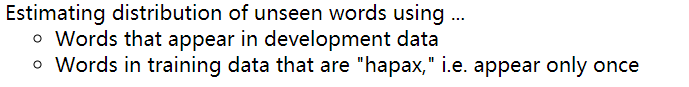
**Smoothing for emission probabilities needs to take these differences into account. Suppose that use Laplace smoothing, separately for each tag, to fill in missing emission probabilities. Then each tag will need a different Laplace constant, i.e. a larger Laplace constant for tags that are more likely to generate unseen words.**

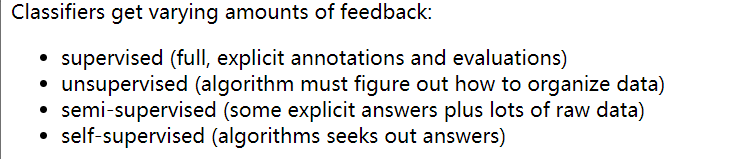


**At what points does an HMM need smoothing?**

near-zero probability of a transition between tags e.g. from Det to Verb

emission probabilities, also need consider open class & closed class

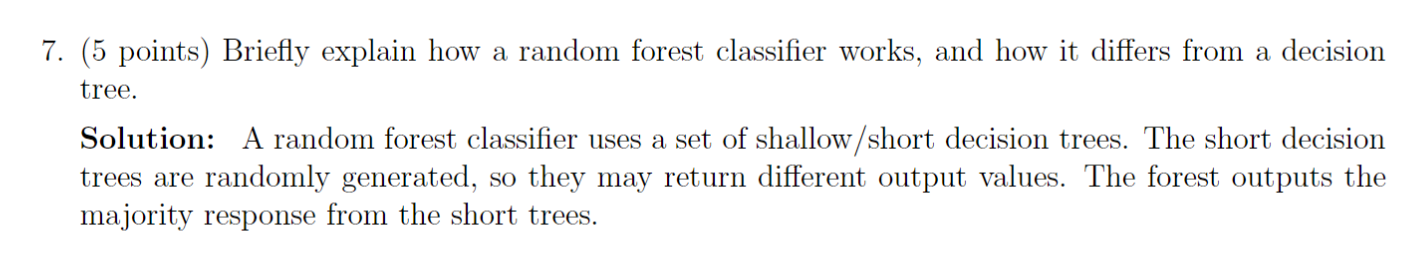




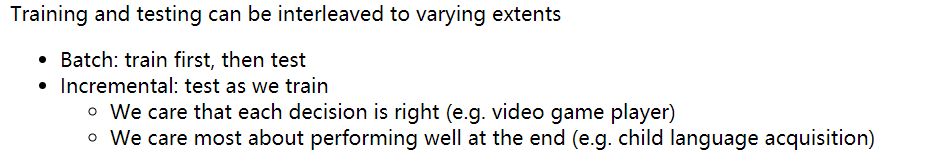
**Decision trees**

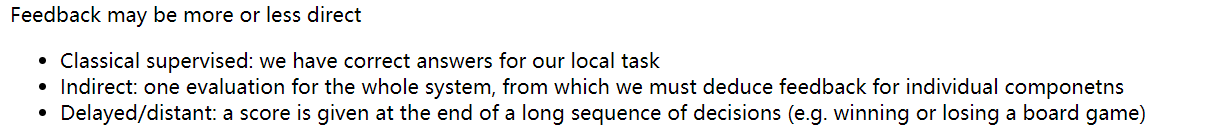
**each node in the tree represents a pool of examples.** Ideally, each leaf node contains only objects of the same type. However, that isn't always true if the classes aren't cleanly distinguished by the available features (see the vowel picture above). So, when we reach a leaf node, the algorithm returns the most frequent label in its pool.

**If we have enough decision levels in our tree, we can force each pool to contain objects of only one type. However, this is not a wise decision. Deep trees are difficult to construct (see the algorithm in Russell and Norvig) and prone to overfitting their training data.** It is more effective to create a "forest" of shallow trees (bounded depth) and have these trees vote on the best label to return. **The set of shallow trees is created by choosing variables and possible split locations randomly.**

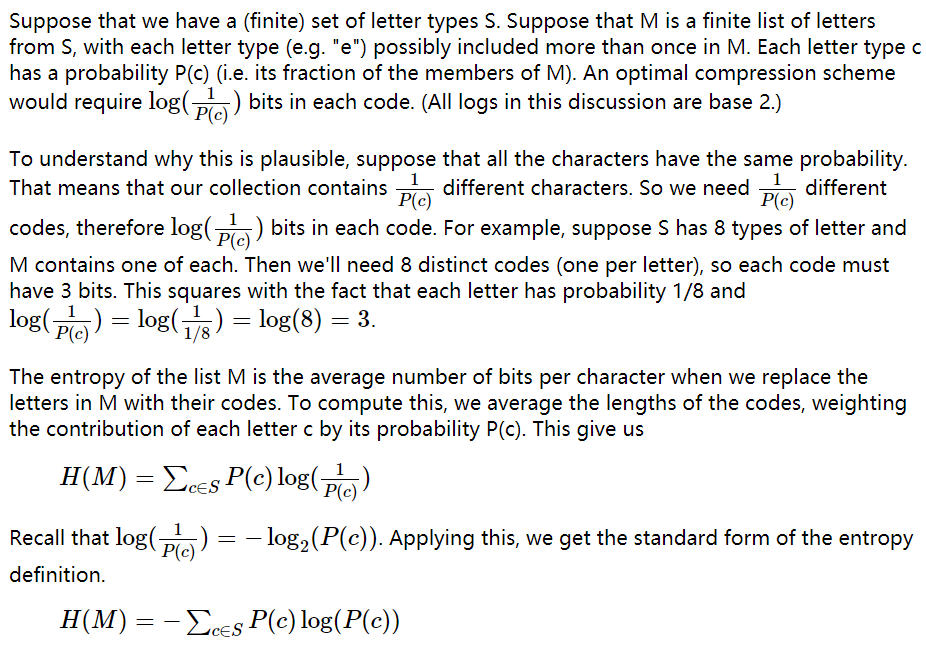


**A good split should produce subpools that are less diverse than the original pool.** If we're choosing between two candidate splits (e.g. two thresholds for the same numerical variable), it's better to pick the one that produces subpools that are more uniform. **Diversity/uniformity is measured using entropy.**





**A good split should produce subpools that are less diverse than the original pool.** If we're choosing between two candidate splits (e.g. two thresholds for the same numerical variable), it's better to pick the one that produces subpools that are more uniform. **Diversity/uniformity is measured using entropy.**

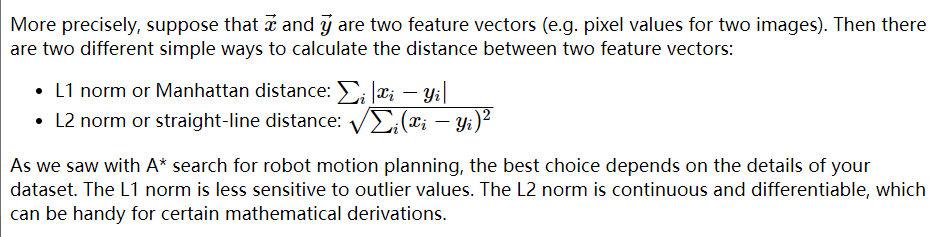


Where P(c) is the possible of letter c occurs in set S

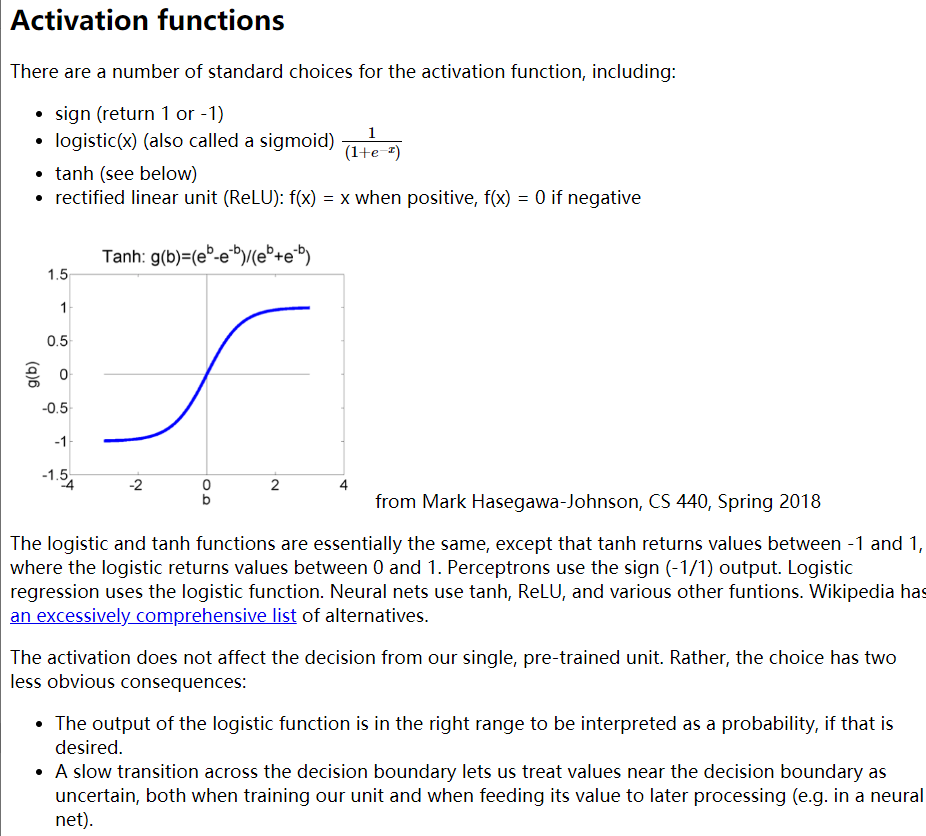
**K nearest neighbors**

Another way to classify an input example is to find the most similar training example and copy its label.

**This "nearest neighbor" method is prone to overfitting, especially if the data is messy and so examples from different classes are somewhat intermixed. However, performance can be greatly improved by finding the k nearest neighbors for some small value of k (e.g. 5-10).**

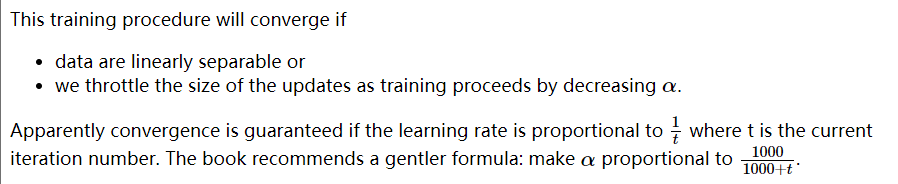


**The L1 norm is less sensitive to outlier values. The L2 norm is continuous and differentiable, which can be handy for certain mathematical derivations.**



**The output of logistic function is in the right range to be interpreted as a probability**

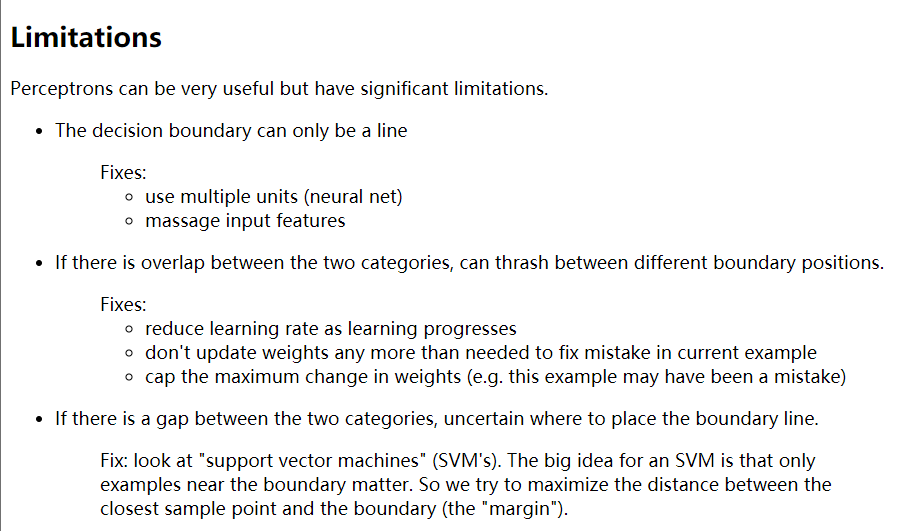
**A slow transition across the decision boundary lets us treat values near the boundary as uncertain.**



**One run through training data through the update rule is called an "epoch."**

**One generally uses multiple epochs, i.e. the algorithm sees each training pair several times.**

Datasets often have strong local correlations, because they have been formed by concatenating sets of data (e.g. documents) that each have a strong topic focus. **So it's often best to process the individual training examples in a random order.**



Multi-class perceptrons

