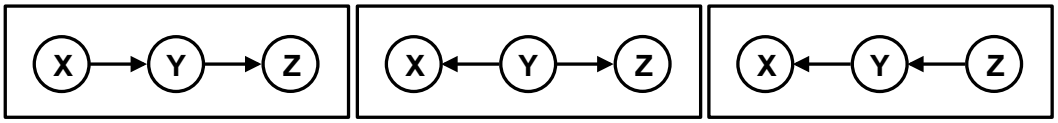
CSE 150. Homework 4 *Summer 2014*

Out: *Fri Aug 15*
Due: *Tue Aug 19 @ 5pm*

4.1 Maximum likelihood estimation

Consider the three DAGs shown below, which are defined over the same nodes X, Y, and Z but differ in the directionality of their edges.



For these DAGs, consider the maximum likelihood CPTs obtained from "fully observed" data $\{(x_t, y_t, z_t)\}_{t=1}^{T}$ in which each example provides a complete instantiation of the nodes X, Y, Z. Also, let $C(x)$ count the number of examples in which $X = x$, let $C(y)$ count the number of examples in which $Y = y$, let $C(z)$ count the number of examples in which $Z = z$, let $C(x, y)$ count the number of examples in which $X = x$ and $Y = y$, and let $C(y, z)$ count the number of examples in which $Y = y$ and $Z = z$.

(a) Express the maximum likelihood estimates for $P(X)$, $P(Y|X)$, and $P(Z|Y)$ in terms of the counts of x, y, and z. Note that these are the CPTs of the left DAG.

(b) Express the maximum likelihood estimates for $P(Y)$, $P(X|Y)$, and $P(Z|Y)$ in terms of the counts of x, y, and z. Note that these are the CPTs of the middle DAG.

(c) Express the maximum likelihood estimates for $P(Z)$, $P(Y|Z)$, and $P(X|Y)$ in terms of the counts of x, y, and z. Note that these are the CPTs of the right DAG.

(d) From your answers in parts (a-c), show that the maximum likelihood CPTs in these different DAGs give rise to the same joint distribution over X, Y, and Z.

(e) Are there any conditional independence relations implied by one DAG that are not implied by the others? Briefly justify your answer.

4.2 Statistical language modeling

In this problem, you will explore some simple statistical models of English text. Download the data files on the course website for this assignment. These files contain unigram and bigram counts for 500 frequently occurring tokens in English text. These tokens include actual words as well as punctuation symbols and other textual markers. In addition, an "unknown" token is used to represent all words that occur outside this basic vocabulary.

(a) Compute the maximum likelihood estimate of the unigram distribution $P_u(w)$ over words w. Print out a table of all the words w that start with the letter "A", along with their unigram probabilities $P_u(w)$. (You do not need to print out the full unigram distribution over all 500 words.)

(b) Compute the maximum likelihood estimate of the bigram distribution $P_b(w'|w)$. Print out a table of the ten most likely words w' to follow the word "THE", along with their bigram probabilities $P_b(w'|w = \mathrm{THE})$. (You do not need need to print out the full bigram matrix.)

(c) Consider the sentence **"The stock market fell by one hundred points last week."** Ignoring punctuation, compute and compare the log-likelihoods (using the natural logarithm) of this sentence under the unigram and bigram models:

$$\begin{aligned}\mathcal{L}_u &= \log\Big[P_u(\textbf{the})\, P_u(\textbf{stock})\, P_u(\textbf{market})\, \ldots\, P_u(\textbf{points})\, P_u(\textbf{last}) P_u(\textbf{week})\Big] \\ \mathcal{L}_b &= \log\Big[P_b(\textbf{the}|\langle\textbf{s}\rangle)\, P_b(\textbf{stock}|\textbf{the})\, P_b(\textbf{market}|\textbf{stock})\, \ldots\, P_b(\textbf{last}|\textbf{points})\, P_b(\textbf{week}|\textbf{last})\Big]\end{aligned}$$

In the equation for the bigram log-likelihood, the token $\langle s\rangle$ is used to mark the beginning of a sentence. Which model yields the highest log-likelihood?

(d) Consider the sentence **"The sixteen officials sold fire insurance."** Ignoring punctuation, compute and compare the log-likelihoods (using the natural logarithm) of this sentence under the unigram and bigram models:

$$\begin{aligned}\mathcal{L}_u &= \log\Big[P_u(\textbf{the})\, P_u(\textbf{sixteen})\, P_u(\textbf{officials})\, \ldots\, P_u(\textbf{sold})\, P_u(\textbf{fire})\, P_u(\textbf{insurance})\Big] \\ \mathcal{L}_b &= \log\Big[P_b(\textbf{the}|\langle\textbf{s}\rangle)\, P_b(\textbf{sixteen}|\textbf{the})\, P_b(\textbf{officials}|\textbf{sixteen})\, \ldots\, P_b(\textbf{fire}|\textbf{sold})\, P_b(\textbf{insurance}|\textbf{fire})\Big]\end{aligned}$$

Which pairs of adjacent words in this sentence are not observed in the training corpus? What effect does this have on the log-likelihood from the bigram model?

(e) Consider the so-called *mixture* model that predicts words from a weighted interpolation of the unigram and bigram models:

$$P_m(w'|w) = (1-\lambda)P_u(w') + \lambda P_b(w'|w),$$

where $\lambda \in [0, 1]$ determines how much weight is attached to each prediction. Under this mixture model, the log-likelihood of the sentence from part (d) is given by:

$$\mathcal{L}_m \;=\; \log\Big[P_m(\textbf{the}|\langle\textbf{s}\rangle)\, P_m(\textbf{sixteen}|\textbf{the})\, P_m(\textbf{officials}|\textbf{sixteen})\;\ldots\; P_m(\textbf{fire}|\textbf{sold})\, P_m(\textbf{insurance}|\textbf{fire})\Big]\,.$$

Compute and plot the value of this log-likelihood $\mathcal{L}_m$ (using the natural logarithm) as a function of the parameter $\lambda \in [0, 1]$. From your results, deduce the optimal value of λ to two significant digits.

Please turn in a printed copy of all your source code for this assignment.
You may program in the language of your choice.
