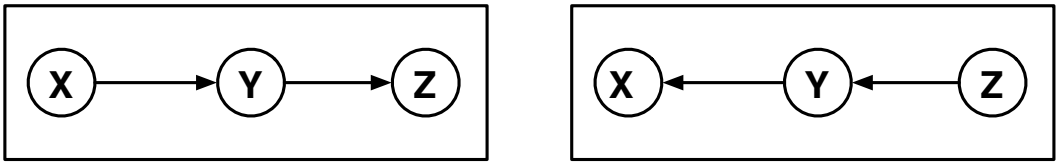
CSE 150. Assignment 4 *Spring 2017*

Out: *Tue May 02*
Due: *Tue May 09*

Recommended: The Unreasonable Effectiveness of Data
`https://www.youtube.com/watch?v=yvDCzhbjYWs`

4.1 Maximum likelihood estimation

Consider the two 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 $C(z)$ 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 $P(Z)$, $P(Y|Z)$, and $P(X|Y)$ in terms of the counts of x, y, and z. Note that these are the CPTs of the right DAG.

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

(d) Are there any conditional independence relations implied by one DAG that are not implied by the other? Briefly justify your answer. Is it consistent with your finding in part (c)?

4.2 Survey

This week you will receive a link to a survey on movies released in 2016; please complete it. We are collecting this data for a future assignment in which you will build a simple movie recommendation system.

4.3 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 "M", 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 "ONE", along with their bigram probabilities $P_b(w'|w = \text{ONE})$. (You do not need need to print out the full bigram matrix.)

(c) Consider the sentence **"The 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:

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

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 fourteen 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:

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

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:

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

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.

(f) **Turn in a printed hard copy of your source code for parts (a) through (e) of this problem. As usual, you may program in the language of your choice.**

4.4 Markov modeling

In this problem, you will construct and compare unigram and bigram models defined over the four-letter alphabet $\mathcal{A} = \{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}\}$. Consider the following 16-token sequence $\mathcal{S}$:

$$\mathcal{S} = \text{“}\texttt{a a b b c c d d d d c c b b a a}\text{”}$$

(a) **Unigram model**

Let τ_ℓ denote the ℓth token of this sequence, and let $L = 16$ denote the total sequence length. The overall likelihood of this sequence under a unigram model is given by:

$$P_U(\mathcal{S}) = \prod_{\ell=1}^{L} P_1(\tau_\ell),$$

where $P_1(\tau)$ is the unigram probability for the token $\tau \in \mathcal{A}$. Compute the maximum likelihood estimates of these unigram probabilities on the training sequence $\mathcal{S}$. Complete the table with your answers.

τ	a	b	c	d
$P_1(\tau)$				

(b) **Bigram model**

Suppose that the overall likelihood of the sequence $\mathcal{S}$ under a bigram model is computed by:

$$P_B(\mathcal{S}) = P_1(\tau_1) \prod_{\ell=2}^{L} P_2(\tau_\ell | \tau_{\ell-1}),$$

where $P_2(\tau'|\tau)$ is the bigram probability that token $\tau \in \mathcal{A}$ is followed by token $\tau' \in \mathcal{A}$. Compute the maximum likelihood estimates of these bigram probabilities on the training sequence $\mathcal{S}$. Complete the table with your answers.

$P_2(\tau'\|\tau)$	τ' = a	τ' = b	τ' = c	τ' = d
τ = a	$\frac{2}{3}$	$\frac{1}{3}$	0	0
τ = b				
τ = c				
τ = d				

(c) **Likelihoods**

Consider again the training sequence $\mathcal{S}$, as well as three test sequences $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ of the same length, shown below. Note that $\mathcal{T}_2$ and $\mathcal{T}_3$ contain bigrams (underlined) that are not in the training sequence $\mathcal{S}$.

$$\begin{aligned}
\mathcal{S} &= \text{“a a b b c c d d d d c c b b a a”} \\
\mathcal{T}_1 &= \text{“d c d c d c d c d c d c d c d c”} \\
\mathcal{T}_2 &= \text{“a a a a d d d d c c c c b b b b”} \\
\mathcal{T}_3 &= \text{“d a d a d a d a d a d a d a d a”}
\end{aligned}$$

Consider the probabilities of these sequences under the unigram and bigram models from parts (a) and (b) of this problem (i.e., the models that you estimated from the training sequence $\mathcal{S}$). For each of the following, indicate whether the probability on the left is equal $(=)$, greater $(>)$, or less $(<)$ than the probability on the right.

Note: you can (and should) answer these questions without explicitly computing the numerical values of the expressions on the left and right hand sides.

$P_U(\mathcal{S})$ ☐ $P_U(\mathcal{T}_1)$

$P_U(\mathcal{S})$ ☐ $P_U(\mathcal{T}_2)$

$P_U(\mathcal{S})$ ☐ $P_U(\mathcal{T}_3)$

$P_B(\mathcal{T}_1)$ ☐ $P_B(\mathcal{S})$

$P_B(\mathcal{T}_2)$ ☐ $P_B(\mathcal{S})$

$P_B(\mathcal{T}_2)$ ☐ $P_B(\mathcal{T}_3)$

$P_U(\mathcal{S})$ ☐ $P_B(\mathcal{S})$

$P_U(\mathcal{T}_1)$ ☐ $P_B(\mathcal{T}_1)$

$P_U(\mathcal{T}_2)$ ☐ $P_B(\mathcal{T}_2)$

$P_U(\mathcal{T}_3)$ ☐ $P_B(\mathcal{T}_3)$

(d) **Likelihoods**

Consider the model obtained by linear interpolation (or mixing) of the unigram and bigram models estimated in part (a) of this problem:

$$P_M(\tau'|\tau) \;=\; (1-\lambda)P_1(\tau') + \lambda P_2(\tau'|\tau),$$

with mixing coefficient $\lambda \in [0,1]$. For a sequence of tokens of length L, the mixture model computes the log-likelihood as:

$$\mathcal{L} \;=\; \log P_1(\tau_1) + \sum_{\ell=2}^{L} \log P_M(\tau_\ell|\tau_{\ell-1}).$$

Naturally, this value varies as a function of the coefficient λ. For λ near zero, it is close to the log-likelihood of the unigram model; for λ near one, it is close to that of the bigram model. This last part of this problem asks you to consider, for each of the sequences below, the *qualitative* behavior of the mixture model's log-likelihood as a function of $\lambda \in [0,1]$. (For instance, is this function constant, or if not, where do its maximum and minimum occur?)

The plots below illustrate four possible behaviors of the mixture model's log-likelihood as a function of $\lambda \in [0,1]$. For each sequence below, indicate the one plot (either A, B, C, or D) that sketches the correct qualitative behavior.

$\mathcal{S}$ = "a a b b c c d d d d c c b b a a" ☐

$\mathcal{T}_1$ = "d c d c d c d c d c d c d c d c" ☐

$\mathcal{T}_2$ = "a a a a d d d d c c c c b b b b" ☐

$\mathcal{T}_3$ = "d a d a d a d a d a d a d a d a" ☐