

Out: Tue May 01

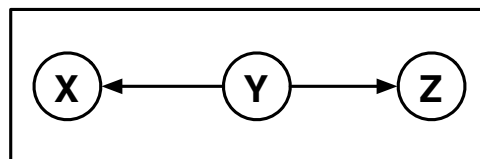
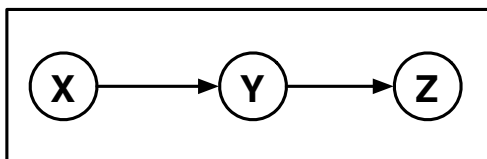
Due: Tue May 08

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.



Handwritten notes showing a table of counts for the data set:

	x	y	z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
...
T	x_T	y_T	z_T

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 $\text{count}(X = x)$, $\text{count}(Y = y)$, $\text{count}(X = x, Y = y)$, and $\text{count}(Y = y, Z = z)$ indicate the number of examples in the data set for which the nodes assume their indicated values.

- (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 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; please complete it. Also, please provide your student ID on this homework assignment so we can check it against the one you provide on the survey. 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 “S”, along with their unigram (normalized) 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 $P_b(w'|w = \text{THE})$. (You do not 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:

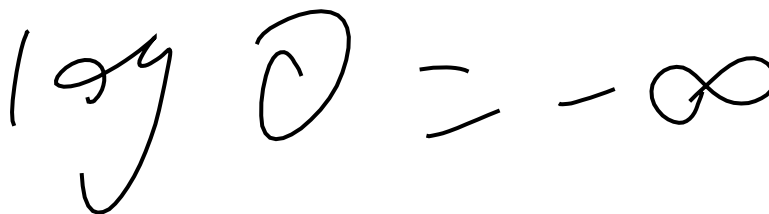
$$\begin{aligned}\mathcal{L}_u &= \log \left[P_u(\text{the}) P_u(\text{stock}) P_u(\text{market}) P_u(\text{fell}) \dots P_u(\text{points}) P_u(\text{last}) P_u(\text{week}) \right] \\ \mathcal{L}_b &= \log \left[P_b(\text{the}|\langle s \rangle) P_b(\text{stock}|\text{the}) P_b(\text{market}|\text{stock}) P_b(\text{fell}|\text{market}) \dots P_b(\text{last}|\text{points}) P_b(\text{week}|\text{last}) \right]\end{aligned}$$

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:

$$\begin{aligned}\mathcal{L}_u &= \log \left[P_u(\text{the}) P_u(\text{sixteen}) P_u(\text{officials}) P_u(\text{sold}) P_u(\text{fire}) P_u(\text{insurance}) \right] \\ \mathcal{L}_b &= \log \left[P_b(\text{the}|\langle s \rangle) P_b(\text{sixteen}|\text{the}) P_b(\text{officials}|\text{sixteen}) \dots P_b(\text{fire}|\text{sold}) P_b(\text{insurance}|\text{fire}) \right]\end{aligned}$$

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?



A handwritten equation in black ink: $\log 0 = -\infty$. The 'log' is written with a large 'l' and 'o', and the '0' is a simple circle. The equals sign is a horizontal line, and the infinity symbol is a circle with a horizontal line through it.

- (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:

$$\mathcal{L}_m = \log \left[P_m(\text{the}|\langle s \rangle) P_m(\text{sixteen}|\text{the}) P_m(\text{officials}|\text{sixteen}) \dots P_m(\text{fire}|\text{sold}) P_m(\text{insurance}|\text{fire}) \right].$$

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 scripts and/or 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 $\mathcal{A} = \{a, b, c, d\}$. Consider the following 16-token sequence \mathcal{S} :

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

(a) Unigram model

Let τ_ℓ denote the ℓ th 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.

$$P_1(a) = \frac{\text{count}(a)}{L}$$

τ	a	b	c	d
$P_1(\tau)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$$= \frac{4}{16} = \frac{1}{4} = P_1(b)$$

(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.

bbbbc

$$P_2(a|b)$$

$$= \frac{\text{count}(X_i=a, P_{i-1}=b)}{\text{count}(P_{i-1}=b)}$$

$$= \frac{0}{4} = 0$$

τ

$P_2(\tau' \tau)$	a	b	c	d
a	$\frac{2}{3}$	0	0	$\frac{1}{3}$
b	0	$\frac{3}{4}$	$\frac{1}{4}$	0
c	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
d	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$

ccb .. ccd

$$P_2(b|c) = \frac{\text{count}(X_i=b, P_{i-1}=c)}{\text{count}(P_{i-1}=c)}$$

$$= \frac{1}{4}$$

ddc .. dda

$$P_2(c|d) = \frac{\text{count}(X_i=c, P_{i-1}=d)}{\text{count}(P_{i-1}=d)} = \frac{1}{4}$$

(c) **Likelihoods**

Consider again the training sequence S , as well as three test sequences T_1 , T_2 , and T_3 of the same length, shown below. Note that T_2 and T_3 contain bigrams (underlined) that are not in the training sequence S .

$S =$ "a a d d c c b b b b c c d d a a"
 $T_1 =$ "b c b c b c b c b c b c b c"
 $T_2 =$ "a a a a b b b b c c c c d d d d"
 $T_3 =$ "b a b a b a b a b a b a b a"

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 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.

$$P_U(S) = P_U(a)^4 \cdot P_U(b)^4 \cdot P_U(c)^4 \cdot P_U(d)^4 = \left(\frac{1}{4}\right)^{16}$$

$$P_U(S) \boxed{=} P_U(T_1)$$

$$P_U(S) \boxed{=} P_U(T_2)$$

$$P_U(S) \boxed{=} P_U(T_3)$$

$$P_U(T_1) = P_U(b) \cdot P_U(c) \cdot \dots \cdot P_U(b) \\ P_U(T_2) = P_U(b)^8 \cdot P_U(c)^8 = \left(\frac{1}{4}\right)^{16}$$

$$P_B(S) = P_U(a) P_B(1a) P_B(d|a) \cdot \dots \cdot P_B(a|d) P_B(a|a) = \frac{1}{4} \cdot \frac{2}{3} \cdot \dots \cdot \frac{1}{4} \cdot \frac{2}{3}$$

$$P_B(S) \boxed{>} P_B(T_1)$$

$$P_B(S) \boxed{>} P_B(T_2)$$

$$P_B(T_2) \boxed{=} P_B(T_3)$$

$$P_B(T_1) = P_U(b) \cdot P_B(c|b)^8 \cdot P_B(b|c)^7 = \left(\frac{1}{4}\right)^{16}$$

$$P_B(T_2) = 0 \text{ since } P_B(b|a) = 0$$

$$P_B(T_3) = 0 \text{ since } P(a|b) = 0$$

$$\left(\frac{1}{4}\right)^{16} < P_B(S) \boxed{>} P_U(S) = \left(\frac{1}{4}\right)^{16}$$

$$\left(\frac{1}{4}\right)^{16} = P_B(T_1) \boxed{=} P_U(T_1) = \left(\frac{1}{4}\right)^{16}$$

$$0 = P_B(T_2) \boxed{<} P_U(T_2) = \left(\frac{1}{4}\right)^{16}$$

$$0 = P_B(T_3) \boxed{<} P_U(T_3) = \left(\frac{1}{4}\right)^{16}$$

$0 \rightarrow 1$
 $x \uparrow \rightarrow$ moving from unigram to bigram

when $\lambda = 0$,

$$P_{12}(\tau' | \tau) = P_1(\tau')$$

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

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.

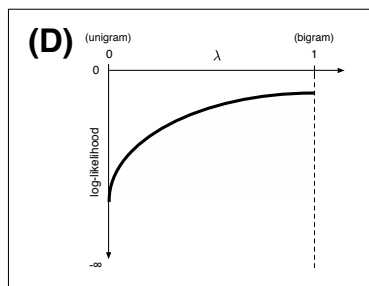
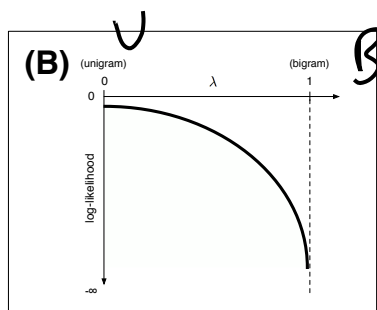
$$p_n(s) \quad p_n(T_1)$$

$P_B(s) > P_U(s)$

$$\boxed{A} \quad P_B(\tau_i) = P_B(\tau_i)$$

$$\left. \begin{matrix} C \\ B \end{matrix} \right\} P_B < P_u$$

B


$$\text{at } \lambda = \frac{1}{2}, \mathcal{L}'(T_2) > \mathcal{L}(T_3)$$