CS 188 Spring 2014 Introduction to Artificial Intelligence Written HW9

INSTRUCTIONS

- **Due:** Monday, April 14th 2014 11:59 PM
- **Policy:** Can be solved in groups (acknowledge collaborators) but must be written up individually. However, we strongly encourage you to first work alone for about 30 minutes total in order to simulate an exam environment. Late homework will not be accepted.
- **Format:** You must solve the questions on this handout (either through a pdf annotator, or by printing, then scanning; we recommend the latter to match exam setting). Alternatively, you can typeset a pdf on your own that has answers appearing in the same space (check edx/piazza for latex templating files and instructions). **Make sure that your answers (typed or handwritten) are within the dedicated regions for each question/part. If you do not follow this format, we may deduct points.**
- **How to submit:** Go to www.pandagrader.com. Log in and click on the class CS188 Spring 2014. Click on the submission titled Written HW 9 and upload your pdf containing your answers. If this is your first time using pandagrader, you will have to set your password before logging in the first time. To do so, click on "Forgot your password" on the login page, and enter your email address on file with the registrar's office (usually your @berkeley.edu email address). You will then receive an email with a link to reset your password.

Last Name	Chen
First Name	Jianzhong
SID	23478230
Email	chenjianzhong@berkeley.edu
Collaborators	None

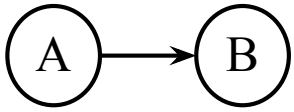
For staff use only

Q. 1	Q. 2	Total
/10	/20	/30

Q1. [10 pts] Machine Learning: Overfitting

Suppose we are doing parameter estimation for a Bayes' net, and have a training set and a test set. We learn the model parameters (probabilities) from the training set and consider how the model applies to the test set. In this question, we will look at how the average log likelihood of the training dataset and the test dataset varies as a function of the number of samples in the training dataset.

Consider the Bayes' net shown on the right. Let $x_1 \dots x_m$ be the m training samples, and $x'_1 \dots x'_n$ be the n test samples, where $x_i = (a_i, b_i)$. Recall that once we have estimated the required model parameters (conditional probability tables) from the training data, the likelihood L for any point x_i is given by:



$$L(x_i) = P(x_i) = P(a_i)P(b_i|a_i)$$

We additionally define the *log-likelihood* of the point x_i, to be $LL(x_i) = \log(L(x_i))$.

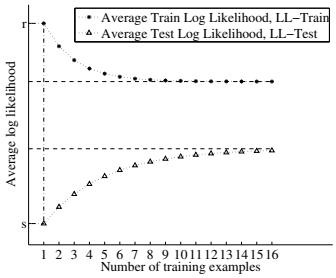
We use the log likelihood for a single point to define the average log likelihood on the training set (LL-Train) and the test set (LL-Test) as follows:

$$\text{LL-Train} = \frac{1}{m}\sum_{i=1}^{m} LL(x_i) = \frac{1}{m}\sum_{i=1}^{m} \log\left(L(x_i)\right) = \frac{1}{m}\sum_{i=1}^{m} \log\left(P(a_i)P(b_i|a_i)\right)$$

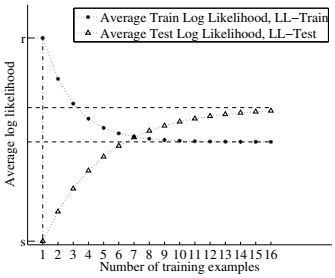
$$\text{LL-Test} = \frac{1}{n}\sum_{i=1}^{n} LL(x'_i) = \frac{1}{n}\sum_{i=1}^{n} \log\left(L(x'_i)\right) = \frac{1}{n}\sum_{i=1}^{n} \log\left(P(a'_i)P(b'_i|a'_i)\right)$$

We assume the test set is very large and fixed, and we will study what happens as the training set grows.

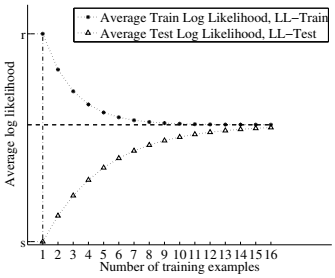
Consider the following graphs depicting the average log likelihood on the Y-axis and the number of training examples on the X-axis.



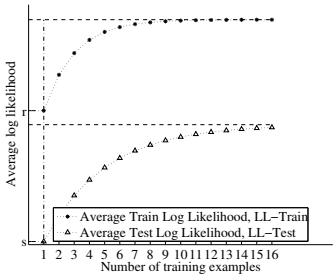
(a) A



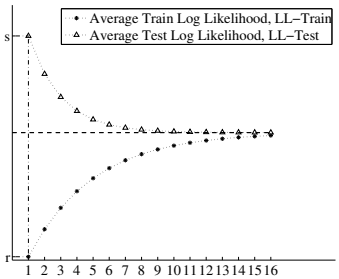
(b) B



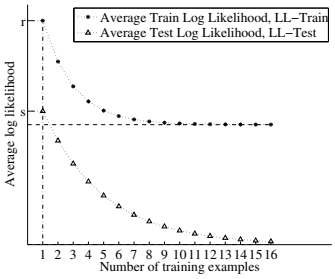
(c) C



(d) D



(e) E



(f) F

(a) [2 pts] Which graph most accurately represents the typical behaviour of the average log likelihood of the training and the testing data as a function of the number of training examples?

○ *A* ○ *B* ● *C* ○ *D* ○ *E* ○ *F*

(b) [2 pts] Suppose our dataset contains exactly one training example. What is the value of LL-Train?

○ $-\infty$ ○ -2 ○ -1 ○ $-1/2$ ● 0

○ $1/2$ ○ 1 ○ 2 ○ ∞

○ *Can* be determined but is not equal to any of the provided options ○ *Cannot* be determined from the information provided

(c) [3 pts] If we did Laplace Smoothing with $k = 5$, how would the values of LL-Train and LL-Test change in the limit of infinite training data? Mark all that are correct. Briefly justify your answer.

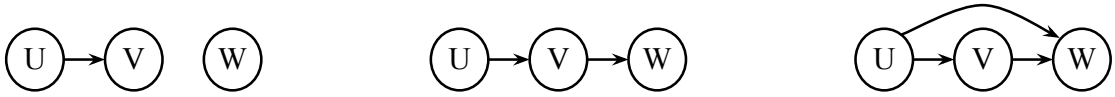
● LL-Train would remain the same. ● LL-Test would remain the same.

○ LL-Train would go up. ○ LL-Test would go up.

○ LL-Train would go down. ○ LL-Test would go down.

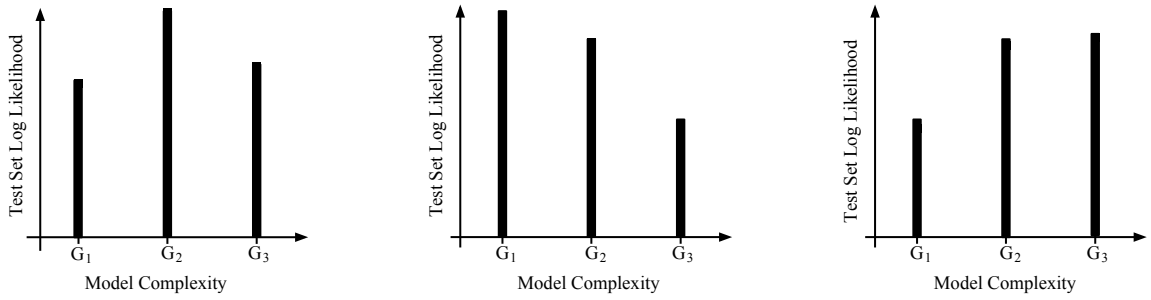
1c Explanation: Since LL-Train and LL-Test are under the limit of infinite data, and Laplace smoothing will only add only a small amount of samples, which is too few to change the value of LL-Train and LL-Test.

(d) [3 pts] Consider the following increasingly complex Bayes' nets: G_1, G_2, and G_3.



G_1 G_2 G_3

Consider the following graphs which plot the test likelihood on the Y-axis for each of the Bayes' nets G_1, G_2, and G_3



For each scenario in the column on the left, select the graph that best matches the scenario. Pick each graph exactly once. Briefly justify your answer.

Small amount of training data: ○ *A* ● *B* ○ *C*

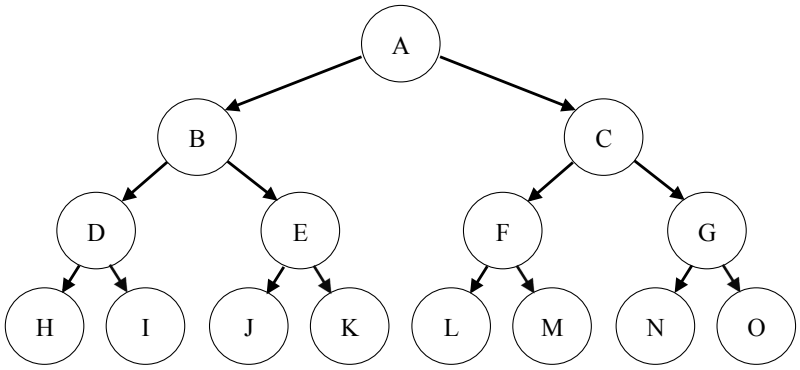
Medium amount of training data: ● *A* ○ *B* ○ *C*

Large amount of training data: ○ *A* ○ *B* ● *C*

1d Explanation: With fewer data, a simple Bayes' net can generalize pretty well, but a complex Bayes' net will overfit. However, for a large amount of data, a simple Bayes' net will underfit, but a complex Bayes' net can simulate the actually distribution.

Q2. [14 pts] Occupy Cal

You are at Occupy Cal, and the leaders of the protest are deciding whether or not to march on California Hall. The decision is made centrally and communicated to the occupiers via the "human microphone"; that is, those who hear the information repeat it so that it propagates outward from the center. This scenario is modeled by the following Bayes net:



A	$P(A)$
$+m$	0.5
$-m$	0.5

$\pi(X)$	X	$P(X\|\pi(X))$
$+m$	$+m$	0.9
$+m$	$-m$	0.1
$-m$	$+m$	0.1
$-m$	$-m$	0.9

Each random variable represents whether a given group of protestors hears instructions to march ($+m$) or not ($-m$). The decision is made at A, and both outcomes are equally likely. The protestors at each node relay what they hear to their two child nodes, but due to the noise, there is some chance that the information will be misheard. Each node except A takes the same value as its parent with probability 0.9, and the opposite value with probability 0.1, as in the conditional probability tables shown.

(a) [2 pts] Compute the probability that node A sent the order to march ($A = +m$) given that both B and C receive the order to march ($B = +m$, $C = +m$).

2a: $p(A = +m|B = +m, C = +m) = \frac{p(A=+m,B=+m,C=+m)}{\sum_a p(A=a,B=+m,C=+m)} = \frac{p(A=+m)p(X=+m|\pi(X)=+m)p(X=+m|\pi(X)=+m)}{\sum_a p(A=a)p(X=+m|\pi(X)=a)p(X=+m|\pi(X)=a)} = \frac{0.5*0.9*0.9}{0.5*0.9*0.9+0.5*0.1*0.1} = \frac{81}{82}$

(b) [2 pts] Compute the probability that D receives the order $+m$ given that A sent the order $+m$.

2b: $p(D = +m|A = +m) = \frac{0.9^2+0.1^2}{0.9^2*0.1^2+2*0.9*0.1} = \frac{41}{50}$

You are at node D, and you know what orders have been heard at node D. Given your orders, you may either decide to march (*march*) or stay put (*stay*). (Note that these actions are distinct from the orders $+m$ or $-m$ that you hear and pass on. The variables in the Bayes net and their conditional distributions still behave exactly as above.) If you decide to take the action corresponding to the decision that was actually made at A (not necessarily corresponding to your orders!), you receive a reward of $+1$, but if you take the opposite action, you receive a reward of -1.

(c) [2 pts] Given that you have received the order $+m$, what is the expected utility of your optimal action? (Hint: your answer to part (b) may come in handy.)

2c: $EU = p(D = +m|A = +m) * (-1) + (1 - p(D = +m|A = +m)) * 1 = -0.64$

Now suppose that you can have your friends text you what orders they have received. (Hint: for the following two parts, you should not need to do much computation due to symmetry properties and intuition.)

(d) [2 pts] Compute the VPI of A given that $D = +m$.

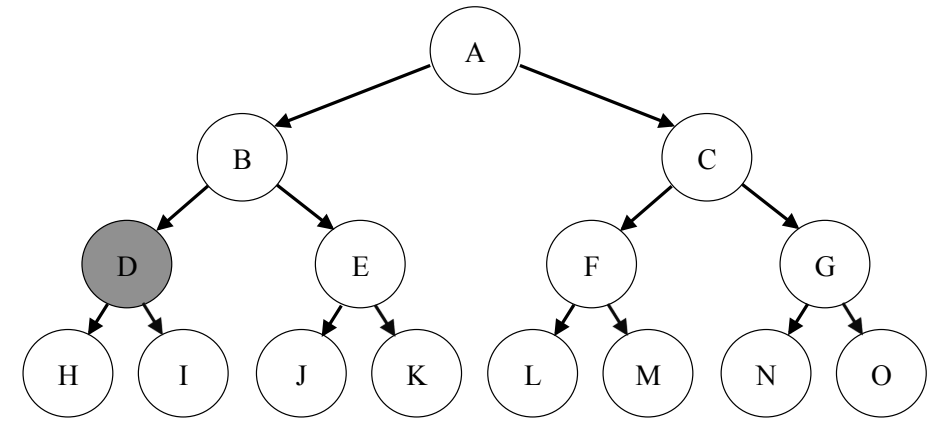
2d: $VPI = 1 - 0.64 = 0.36$

(e) [2 pts] Compute the VPI of F given that $D = +m$.

2e: Since D is independent from F given A, so the VPI of F given D=+m is 0.

For the following parts, you should select nodes in the accompanying diagrams that have the given properties. Use the quantities you have already computed and intuition to answer the following question parts; you should not need to do any computation.

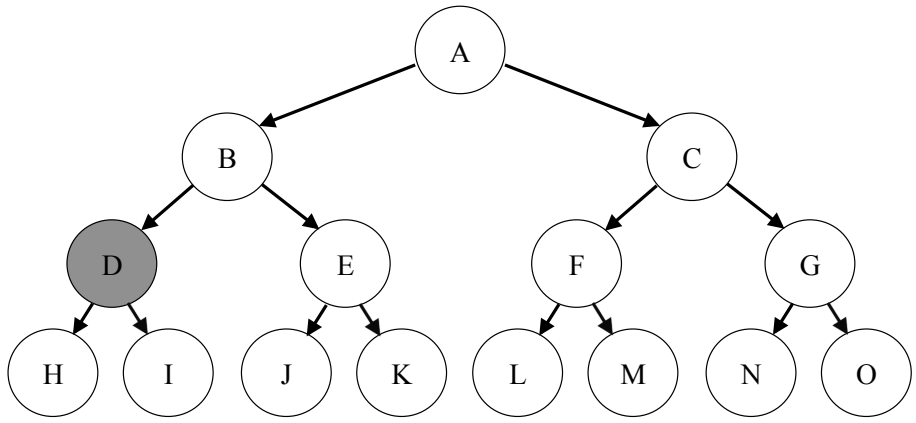
(f) [5 pts] Select the nodes for which knowing the value of that node changes your belief about the decision made at A given evidence at D (i.e. nodes X such that $P(A|X, D) \neq P(A|D)$).



● A	● B	○ C	○ E	
○ F	○ G	○ H	○ I	○ J
○ K	○ L	○ M	○ N	○ O

2f Explanation: Because D is dependent on B, and B is dependent on A.

(g) [5 pts] Select the nodes which have nonzero VPI given evidence at D.



- [x] A
- [x] B
- [ ] C
- [ ] E
- [ ] F
- [ ] G
- [ ] H
- [ ] I
- [ ] J
- [ ] K
- [ ] L
- [ ] M
- [ ] N
- [ ] O

2g Explanation: Same reason as F