Machine Learning (10-701)		Fall 2008
	Final Exam	
Professor: Eric Xing		*Date: December 8, 2008*

. There are 9 questions in this exam (18 pages including this cover sheet)

. Questions are not equally difficult.

. This exam is open to book and notes. Computers, PDAs, Cell phones are not allowed.

. You have three hours.

. Good luck!

Last Name:			
First Name:			
Andrew ID:			
Q	Topic	Max. Score	Score
1	**Assorted Questions**	**20**	
2	**SVM**	**10**	
3	**PCA**	**10**	
4	**Linear Regression**	**12**	
5	**Sampling**	**8**	
6	**EM**	**10**	
7	**Learning Theory**	**10**	
8	**Hidden Markov Models**	**10**	
9	**Bayesian Networks**	**10**	
Total		100	

1 Assorted Questions [20 points]

1. (**True or False**, 2 pts) PCA and Spectral Clustering (such as Andrew Ng's) perform eigen-decomposition on two different matrices. However, the size of these two matrices are the same.

   Solutions: F

2. (**True or False**, 2 pts) The dimensionality of the feature map generated by polynomial kernel (e.g., $K(x, y) = (1 + x \cdot y)^d$)is polynomial wrt the power d of the polynomial kernel.

   Solutions: T

3. (**True or False**, 2 pts) Since classification is a special case of regression, logistic regression is a special case of linear regression.

   Solutions: F

4. (**True or False**, 2 pts) For any two variables x and y having joint distribution $p(x, y)$, we always have $H[x, y] \geq H[x] + H[y]$ where H is entropy function.

   Solutions: F

5. (**True or False**, 2 pts) The Markov Blanket of a node x in a graph with vertex set X is the smallest set Z such that $x \perp X/\{Z \cup x\}|Z$

   Solutions: T

6. (**True or False**, 2 pts) For some directed graphs, moralization decreases the number of edges present in the graph.

   Solutions: F

7. (**True or False**, 2 pts) The L_2 penalty in a ridge regression is equivalent to a Laplace prior on the weights.

   Solutions: F

8. (**True or False**, 2 pts) There is *at least one* set of 4 points in $\Re^3$ that can be shattered by

the hypothesis set of all 2D planes in $\Re^3$.

Solutions: T

9. (**True or False**, 2 pts) The log-likelihood of the data will *always* increase through successive iterations of the expectation maximation algorithm.

   Solutions: F

10. (**True or False**, 2 pts) One disadvantage of Q-learning is that it can only be used when the learner has prior knowledge of how its actions affect its environment.

    Solutions: F

2 Support Vector Machine(SVM) [10 pts]

1. **Properties of Kernel**

   1.1. (2 pts) Prove that the kernel $K(x_1, x_2)$ is symmetric, where x_i and x_j are the feature vectors for i^{th} and j^{th} examples.

   hints: Your proof will not be longer than 2 or 3 lines.

   Solutions: Let $\Phi(x_1)$ and $\Phi(x_2)$ be the feature maps for x_i and x_j, respectively. Then, we have $K(x_1, x_2) = \Phi(x_1)'\Phi(x_2) = \Phi(x_2)'\Phi(x_1) = K(x_2, x_1)$

   1.2. (4 pts) Given n training examples $(x_i, x_j)(i, j = 1, ..., n)$, the kernel matrix $\mathbf{A}$ is an $n \times n$ square matrix, where $\mathbf{A}(i, j) = K(x_i, x_j)$. Prove that the kernel matrix $\mathbf{A}$ is semi-positive definite.

   hints: (1) Remember that an $n \times n$ matrix $\mathbf{A}$ is semi-positive definite iff. for any n dimensional vector $\mathbf{f}$, we have $\mathbf{f}'\mathbf{A}\mathbf{f} \geq 0$. (2) For simplicity, you can prove this statement just for the following particular kernel function: $K(x_i, x_j) = (1 + x_i x_j)^2$.

   Solutions: Let $\Phi(x_i)$ be the feature map for the i^{th} example and define the matrix $\mathbf{B} = [\mathbf{\Phi(x_1)}, ..., \mathbf{\Phi(x_n)}]$. It is easy to verify that $\mathbf{A} = \mathbf{B}'\mathbf{B}$. Then, we have $\mathbf{f'Af} = \mathbf{(Bf)'Bf} = \|\mathbf{Bf}\|^\mathbf{2} \geq \mathbf{0}$

2. **Soft-Margin Linear SVM**. Given the following dataset in 1-d space (Figure 1), which consists of 4 positive data points $\{0, 1, 2, 3\}$ and 3 negative data points $\{-3, -2, -1\}$. Suppose that we want to learn a soft-margin linear SVM for this data set. Remember that the soft-margin linear SVM can be formalized as the following constrained quadratic optimization problem. In this formulation, C is the regularization parameter, which balances the size of margin (i.e., smaller $w^t w$) vs. the violation of the margin (i.e., smaller $\sum_{i=1}^{m} \epsilon_i$).

$$
\begin{aligned}
&\text{argmin}_{\{w,b\}} \ \frac{1}{2} w^t w + C \sum_{i=1}^{m} \epsilon_i \\
&\text{Subject to :} y_i(w^t x_i + b) \geq 1 - \epsilon_i \\
&\epsilon_i \geq 0 \ \forall i
\end{aligned}
$$

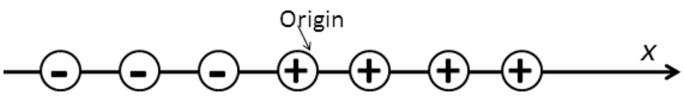


Figure 1: Dataset

2.1 (2 pts) if $C = 0$, which means that we only care the size of the margin, how many support vectors do we have?

Solutions: 7

2.2 (2 pts) if $C \to \infty$, which means that we only care the violation of the margin, how many support vectors do we have?

Solutions: 2

3 Principle Component Analysis (PCA) [10 pts]

1.1 (3 pts) **Basic PCA**

Given 3 data points in 2-d space, $(1,1)$, $(2,2)$ and $(3,3)$,

(a) (1 pt) what is the first principle component?

Solutions: $pc = (1/\sqrt{2}, 1/\sqrt{2})' = (0.707, 0.707)'$, (the negation is also correct)

(b) (1 pt) If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data?

Solutions: $4/3 = 1.33$

(c) (1 pt) For the projected data in (b), now if we represent them in the original 2-d space, what is the reconstruction error?

Solutions: 0

1.2 (7 pts) **PCA and SVD**

Given 6 data points in 5-d space, $(1,1,1,0,0)$, $(-3,-3,-3,0,0)$, $(2,2,2,0,0)$, $(0,0,0,-1,-1)$, $(0,0,0,2,2)$, $(0,0,0,-1,-1)$. We can represent these data points by a 6×5 matrix X, where each row corresponds to a data point:

$$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ -3 & -3 & -3 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix}$$

(a) (1 pt) What is the sample mean of the data set?

Solutions: $[0, 0, 0, 0, 0]$

(b) (3 pts) What is SVD of the data matrix X you choose?

hints: The SVD for this matrix must take the following form, where $a, b, c, d, \sigma_1, \sigma_2$ are the parameters you need to decide.

$$X = \begin{bmatrix} a & 0 \\ -3a & 0 \\ 2a & 0 \\ 0 & b \\ 0 & -2b \\ 0 & b \end{bmatrix} \times \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \times \begin{bmatrix} c & c & c & 0 & 0 \\ 0 & 0 & 0 & d & d \end{bmatrix}$$

Solutions: $a = \pm 1/\sqrt{14} = \pm 0.267$, $b = \pm 1/\sqrt{6} = \pm 0.408$,
$\sigma_1 = 1/(a \cdot c) = \sqrt{42} = 6.48$, $\sigma_2 = 1/(b \cdot d) = \sqrt{12} = 3.46$,
$c = \pm 1/\sqrt{3} = \pm 0.577$,$d = \pm 1/\sqrt{2} = \pm 0.707$.

(c) (1 pt) What is first principle component for the original data points?

Solutions: pc=$\pm[c, c, c, 0, 0] = \pm[0.577, 0.577, 0.577, 0, 0]$ (Intuition: First, we want to notice that the first three data points are co-linear, and so do the last three data points. And also the first three data points are orthogonal to the rest three data points. Then, we want notice that the norm of the first three are much bigger than the last three, therefor, the first pc has the same direction as the first three data points)

(d) (1 pt) If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data?

Solutions: var=$\sigma_1^2/6 = 7$ (Intuition: we just the keep the first three data points, and set the rest three data points as $[0, 0, 0, 0, 0]$ (since they are orthogonal to pc), and then compute the variance among them)

(e) (1 pt) For the projected data in (d), now if we represent them in the original 5-d space, what is the reconstruction error?

Solutions: var=$\sigma_2^2/6 = 2$[1](Intuition, since the first three data points are orthogonal with the rest three, here the rerr is the just the sum of the norm of the last three data points (2+8+2=12), and then divided by the total number (6) of data points, if we use average definition

[1]if you give an answer var=$\sigma_2^2 = 12$, that is also correct. In this case, the reconstruction error is just the sum (not average) among all the data points, which is the definition used in Carlos' lecture notes. But in Bishop's book, he uses the average definition.

4 Linear Regression [12 Points]

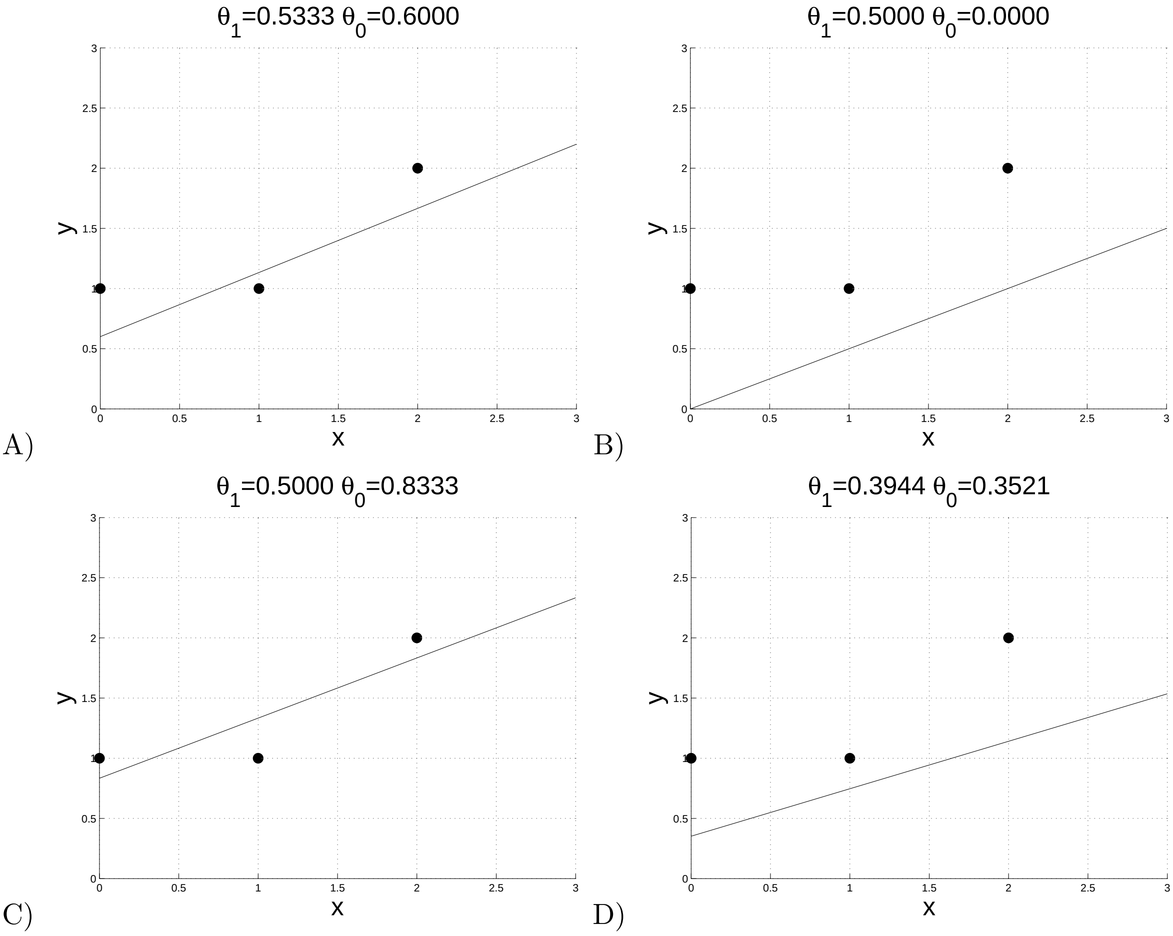


Figure 2: Plots of linear regression results with various regularization

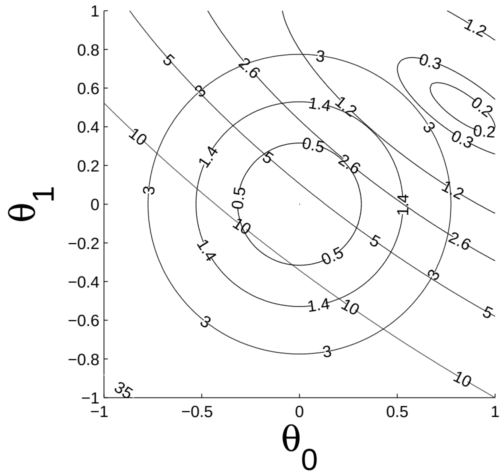
Background: In this problem we are working on linear regression with regularization on points in a 2-D space. Figure 2 plots linear regression results on the basis of three data points, (0,1), (1,1) and (2,2), with different regularization penalties.

As we all know, solving a linear regression problem is about to solve a minimization problem. That is to compute

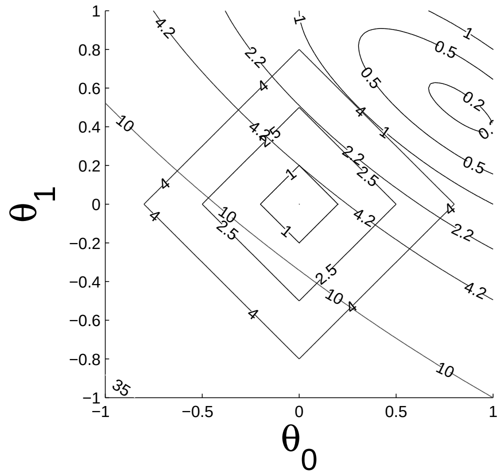
$$\arg\min_{\theta_0,\theta_1} \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 + R(\theta_0, \theta_1)$$

where R represents a regularization penalty which could be L-1 or L-2. In this problem, $n = 3$, $(x_1, y_1) = (0, 1)$, $(x_2, y_2) = (1, 1)$, and $(x_3, y_3) = (2, 2)$. $R(\theta_0, \theta_1)$ could either be $\lambda(|\theta_1| + |\theta_0|)$ or $\lambda(\theta_1^2 + \theta_0^2)$.

However, in stead of computing the derivatives to get a minimum value, we could adopt a geometric method. In this way, rather than letting the square error term and the regularization penalty term vary simultaneously as a function of θ_0 and θ_1, we can fix one and only let the other vary at a time. Having a upper-bound, r, on the penalty, we can replace $R(\theta_0, \theta_1)$ by r, and solve a minimization



(a)



(b)

Figure 3: Contour plots of the decomposition for the linear regression problem with (a) L-2 regularization or (b) L-1 regularization where the ellipsis correspond to the square error term, and circles/squares correspond to the regularization penalty term.

problem on the square error term for any non-negative value of r. Finally, we get the minimum value by enumerating over all possible value of r. That is,

$$\min_{\theta_0,\theta_1} \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 + R(\theta_0,\theta_1) = \min_{r \geq 0} \left\{ \min_{\theta_0,\theta_1} \left\{ \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 \mid R(\theta_0,\theta_1) \leq r \right\} + r \right\}$$

. The value of (θ_0, θ_1) corresponding to the minimum value of the object function can be got at the same time.

In Figure 3, we plot the square error term, $\sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2$, by ellipse contours. The circle contours in Fig 3(a) plots a L-2 penalty with $\lambda = 5$, whereas the square contours in Fig 3(b) plots a L-1 penalty with $\lambda = 5$.

To further explain how it works, the solution to

$$\min_{\theta_0,\theta_1} \left\{ \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 \mid R(\theta_0,\theta_1) \leq r \right\}$$

is the height of the smallest ellipse contour that is tangent with (or contained in) the contour that depict $R(\theta_0, \theta_1) = r$. The desired (θ_0, θ_1) are the coordinates of the tangent point.

Question:

1. Please assign each plot in Figure 2 to one (and only one) of the following regularization methods. You can get some help from Figure 3. Please answer A, B, C or D.

   (a) (2 pts) No regularization (or regularization parameter equals to 0).

$$\sum_{i=1}^{3} (y_i - \theta_1 x_i - \theta_0)^2$$

Solution: C

(b) (3 pts) L-2 regularization with λ being 5.

$$\sum_{i=1}^{3}(y_i - \theta_1 x_i - \theta_0)^2 + \lambda(\theta_1^2 + \theta_0^2) \text{ where } \lambda = 5$$

Solution: D

(c) (3 pts) L-1 regularization with λ being 5.

$$\sum_{i=1}^{3}(y_i - \theta_1 x_i - \theta_0)^2 + \lambda(|\theta_1| + |\theta_0|) \text{ where } \lambda = 5$$

Solution: B

(d) (2 pts) L-2 regularization with λ being 1.

$$\sum_{i=1}^{3}(y_i - \theta_1 x_i - \theta_0)^2 + \lambda(\theta_1^2 + \theta_0^2) \text{ where } \lambda = 1$$

Solution: A

2. (2 pts) If we have much more features (that is more x_i's) and we want to perform feature selection while solving the LR problem, which kind of regularization method do we want to use? (Hint: L-1 or L-2? What about λ?)

Solution: We will choose L-1, and we will use bigger λ when we want fewer effective features.

5 Sampling [8 Points]

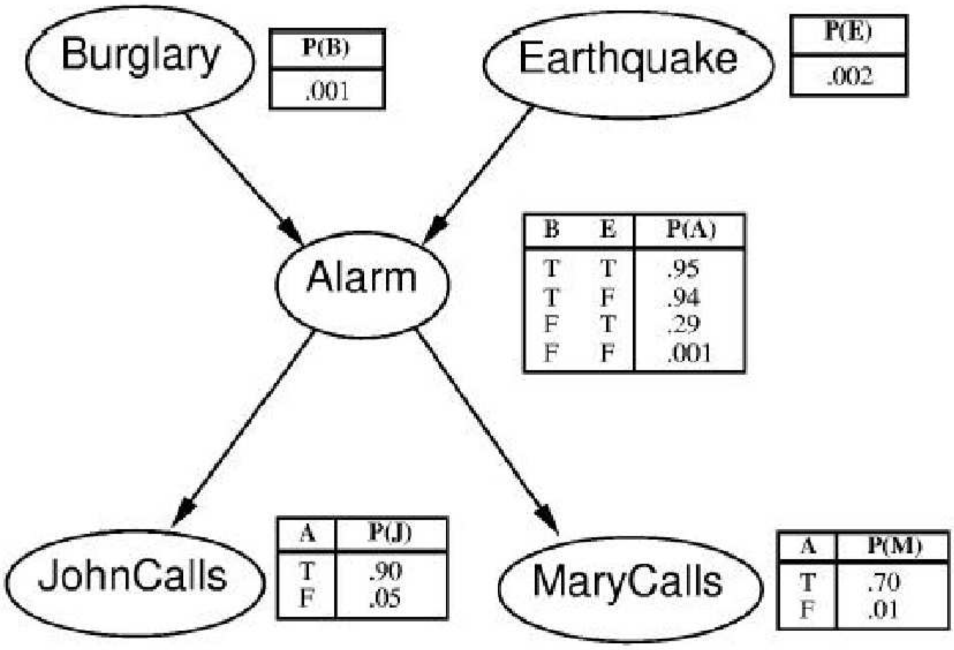


Figure 4: A Bayesian Network for studying sampling

1. (2 pts) Suppose we want to compute $P(B1|E1)$ using the naive sampling method. We draw 1,000,000 sample records in total. How many useful samples do we expect to see? (Hint: $B1$ means Burglary is true.)

   Solution: Records with $E = 1$ are useful samples. So, $1000000 * 0.002 = 2000$.

2. (1 pts) Suppose we want to compute $P(B1|J1)$ using the Gibbs sampling algorithm. How many different states of (B,E,A,J,M) will we observe during the process?

   Solution: There are four variables (B,E,A,M), each of which has two states, so we can observe $2^4 = 16$ different states.

3. (3 pts) Suppose we want to compute $P(B1|J1)$ using the Gibbs sampling algorithm, and we start with state (B1,E0,A0,J1,M0). We choose variable E in the first step. What are the possible states after the first step, and what are their probability of occurrence respectively?

   Solution: The next possible states are (B1,E0,A0,J1,M0) and (B1,E1,A0,J1,M0), because only E may change.

$$P(E1|B1, A0) = \frac{P(E1, B1, A0)}{P(B1, A0)} = \frac{P(E1, B1, A0)}{P(E1, B1, A0) + P(E0, B1, A0)}$$

$$P(E1, B1, A0) = P(E1) * P(B1) * P(A0|E1, B1) = 10^{-7}$$

$$P(E0, B1, A0) = P(E0) * P(B1) * P(A0|E0, B1) = 5.988 * 10^{-5}$$

So, $P(E1|B1, A0) = 0.0017$

$$P(E0|B1, A0) = 1 - P(E1|B1, A0) = 0.9983$$

With probability 0.9983 it will become (B1,E0,A0,J1,M0), and with probability 0.0017 it will become (B1,E1,A0,J1,M0).

4. (2 pts) In Markov Chain Monte Carlo (MCMC), is choosing the transition probabilities to satisfy the property of detailed balance a necessary condition for ensuring that a stationary distribution exists? Please answer Yes or No.

   Solution: No. It is a sufficient condition.

6 Expectation Maximization [10 Points]

Imagine a machine learning class where the probability that a student gets an "A" grade is $\mathbb{P}(A) = 1/2$, a "B" grade $\mathbb{P}(B) = \mu$, a "C" grade $\mathbb{P}(C) = 2\mu$, and a "D" grade $\mathbb{P}(D) = 1/2 - 3\mu$. We are told that c students get a "C" and d students get a "D". We don't know how many students got exactly an "A" or exactly a "B". But we do know that h students got either an a or b. Therefore, a and b are unknown values where $a + b = h$. Our goal is to use expectation maximization to obtain a maximum likelihood estimate of μ.

1. (4 pts) Expectation step: Which formulas compute the expected values of a and b given μ? Circle your answers.

$$\widehat{a} = \frac{1/2}{1/2 + h}\mu \qquad \widehat{b} = \frac{\mu}{1/2 + h}\mu$$

$$**** \, \widehat{a} = \frac{1/2}{1/2 + \mu}h \qquad \widehat{b} = **** \, \frac{\mu}{1/2 + \mu}h$$

$$\widehat{a} = \frac{\mu}{1/2 + \mu}h \qquad \widehat{b} = \frac{1/2}{1/2 + \mu}h$$

$$\widehat{a} = \frac{1/2}{1 + \mu^2}h \qquad \widehat{b} = \frac{\mu}{1 + \mu^2}h$$

Solution: Marked with ****

2. (4 pts) Maximization step: Given the expected values of a and b which formula computes the maximum likelihood estimate of μ? Circle your answer. *Hint:* Compute the MLE of μ assuming unobserved variables are replaced by their expectation.

$$**** \, \widehat{\mu} = \frac{h - a + c}{6(h - a + c + d)}$$

$$\widehat{\mu} = \frac{h - a + d}{6(h - 2a - d)}$$

$$\widehat{\mu} = \frac{h - a}{6(h - 2a + c)}$$

$$\widehat{\mu} = \frac{2(h - a)}{3(h - a + c + d)}$$

Solution: Marked with ****

3. (True/False, 2 pts) Iterating between the E-step and M-step will *always* converge to a local optimum of μ (which may or may not also be a global optimum)? Explain in 1-2 sentences.
Solution: True, the lower bound increases on each iteration.

7 VC-Dimension and Learning Theory [10 Points]

1. (True/False, 2 pts) Can the set of all rectangles in the 2D plane (which includes non axis-aligned rectangles) shatter a set of 5 points? Explain in 1-2 sentences.

   Solution: True, can shatter 5 points along a circle.

2. (2 pts) What is the VC-dimension of k-Nearest Neighbour classifier when $k = 1$? Explain in 1-2 sentences.

   Solution: Infinity since it can shatter an arbitrary training set.

3. (2 pts) Consider the classifier $f(a) = 1$ if $a > 0$ and $f(a) = 0$ otherwise. What is the VC-dimension of $f(sin(\alpha x))$ when α is an adjustable parameter? Explain in 1-2 sentences.

   Solution: Infinity since it can shatter an arbitrary training set.

Consider the following formulas that bound the number of training examples necessary for successful learning:

$$
\begin{aligned}
m &\geq \frac{1}{\epsilon}(\ln(1/\delta) + \ln|H|) \\
m &\geq \frac{1}{2\epsilon^2}(\ln(1/\delta) + \ln|H|) \\
m &\geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))
\end{aligned}
$$

For each of the below questions, **pick the formula** you would use to estimate the number of examples you would need to learn the concept. You do not need to do any computation or plug in any numbers. Explain your answer.

1. (2 pts) Consider instances with two Boolean variables $\{X_1, X_2\}$, and responses Y are given by the XOR function. We try to learn the function $f : X \to Y$ using a *2-layer neural network*.

   Solution: Eq (3) since the hypothesis space is infinite.

2. (2 pts) Consider instances with two Boolean variables $\{X_1, X_2\}$, and responses Y are given by the XOR function. We try to learn the function $f : X \to Y$ using a *depth-two decision tree*. This tree has four leaves, all distance two from the top.

   Solution: Eq(1) because the hyptothesis space is finite and $Y \in H$

8 Hidden Markov Models with continuous emissions (10 points)

In this question, we will study hidden markov models with continuous emissions. We will use the notation used in class, with x^i denoting the output at time i, and y_i denoting the corresponding hidden state. The HMM has K states $\{1 \dots K\}$. The output for state k is obtained by sampling a Gaussian distribution parameterized by mean μ_k and standard deviation σ_k. Thus, we can write the emission probabilitye as $p(x_i|y_i = k, \theta) = \mathcal{N}(x_i|\mu_k, \sigma_k)$. θ is the set of parameters of the HMM, which includes the initial probabilities π, transition probability matrix A and the means and standard deviations $\{\mu_1, \dots, \mu_K, \sigma_1, \dots, \sigma_k\}$.

8.1 Log-likelihood (1 point)

Write down the log-likelihood for a sequence of observations of the emissions $\{x_1, \dots, x_n\}$ when the states (also observed) are $\{y_1, \dots, y_n\}$.

Solution:

$$\log p(x_1, \dots, x_n|y_1, \dots, y_n) = \log \prod_i p(x_i|y_i) \quad (1)$$

$$= \sum_i \log(\mathcal{N}(x_i|\mu_{y_i}, \sigma_{y_i})) \quad (2)$$

8.2 Forward and backward updates (2 points)

Write the forward and backward update equations for this HMM. Explain in a single line how they are different from the updates we studied in class.

Solution:

$$\alpha_t^k = \mathcal{N}(x_t|\mu_k, \sigma_k) \sum_i \alpha_{t-1}^i a_{i,k} \quad (3)$$

$$\beta_t^k = \sum_i a_{k,i} \beta_{t+1}^i \mathcal{N}(x_t|\mu_i, \sigma_i) \quad (4)$$

The equations are similar in form. But in this case, the output probabilities are gaussian rather than multinomial. The outputs are also continuous rather than discrete.

8.3 Supervised parameter learning

We are given a sequence of observations $X = \{x_1, \dots, x_n\}$ and the corresponding hidden states $Y = \{y_1, \dots, y_n\}$. We want to find the parameters θ for the HMM.

1. Are the update equations for A_{ij} and π_i different from the ones obtained for the HMM we studied in class? Explain why or why not (2 points).

   Solution: The update equations for A_{ij} and π_i are the same. They involve only the state transition counts and so are independent of the form chosen for emission probabilities.

2. What are the update equations for the Gaussian parameters μ_k and σ_k ? (Hint: You do not need to derive them. Given the hidden states, the outputs are all independent of each other, and each is sampled from one out of K gaussians.) (2 points)

   Solution:

$$\mu_k = \frac{\sum_i \mathcal{I}[y_i = k] x_i}{\sum_i \mathcal{I}[y_i = k]} \tag{5}$$

$$\sigma_k^2 = \frac{\sum_i \mathcal{I}[y_i = k](x_i - \mu_k)^2}{\sum_i \mathcal{I}[y_i = k]} \tag{6}$$

8.4 Unsupervised parameter learning

Now, we are only given a sequence of observations $X = \{x_1, \ldots, x_n\}$. We want to find the parameters θ for the HMM. (Slide 47 and 48 for the HMM lecture describe the unsupervised learning algorithm for the HMM discussed in class)

8.5 Objective function

The unsupervised learning algorithm optimizes the expected complete log-likelihood. Why is that a reasonable choice for the objective function? (1 point)

Solution: The expected complete log-likelihood is a lower bound to the complete log-likelihood. It is guaranteed to converge to a local optimum of the complete likelihood. Hence it is a reasonable choice for the objective function.

8.5.1 Expected complete LL

What is the expected complete log-likelihood ($\langle l_c(\theta; x, y) \rangle$) for the HMM with continuous gaussian emissions? Just write the expression, a derivation is not necessary.(1 point)

Solution:

$$\langle l_c(\theta; x, y) \rangle = \sum_n \left(\langle y_{n,1}^i \rangle \log \pi_i \right) + \sum_n \sum_{t=2}^{T} \left(\left\langle y_{n,t-1}^i y_{n,t}^j \right\rangle \log a_{i,j} \right) + \sum_n \sum_{t=1}^{T} \left(\langle y_{n,t}^i \rangle \log \mathcal{N}(x_n, t | \mu_i, \sigma_i) \right) \tag{7}$$

8.5.2 Gaussian Parameter estimation

Suppose you want to find ML estimates $\hat{\mu}_k$ and $\hat{\sigma}_k$ for parameters μ_k and σ_k . Will the ML expressions have the same form as those obtained for the means and variances in a mixture of gaussians? Explain in one line. (Hint: Write down the terms in $\langle l_c(\theta; x, y) \rangle$ that are relevant to the optimization (i.e, contain μ_k and σ_k))(1 point)

Solution: Yes, the ML expressions will have same form. The relevant term in $\langle l_c(\theta; x, y) \rangle$ is only the last term, which closely resembles the expected complete log-likelihood for a gaussian mixture model. In this case the $p(y=1|x)$ term is computed using the forward backward algorithm rather than by simply using Bayes rule (as is done for a mixture of gaussians).

9 Bayesian Networks (10 points)

Consider the Bayesian network shown in Figure 5. All the variables are boolean.

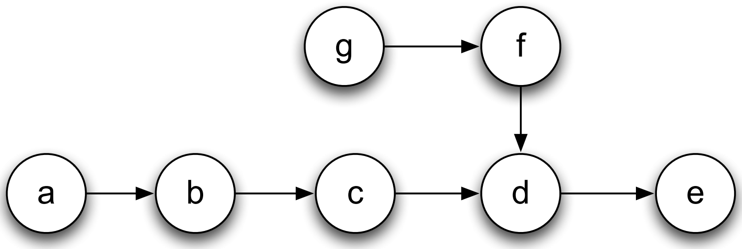


Figure 5: Bayesian network for Question 9.2 and 9.3

9.1 Likelihood

Write the expression for the joint likelihood of the network in its factored form. (2 points)

Solution: $p(a,b,c,d,e,f,g) = p(a)p(b|a)p(c|b)p(d|c,f)p(e|d)p(f|g)p(g)$

9.2 D-separation

1. Let $X = \{c\}, Y = \{b,d\}, Z = \{a,e,f,g\}$. Is $X \perp Z|Y$? If yes, explain why. If no, show a path from X to Z that is not blocked. (2 points)

   Solution: No, $X \not\perp Z|Y$. The path $c \to d \to f$ is not blocked since the v-structure at d is observed.

2. Suppose you are allowed to choose a set W such that $W \subset Z$. Then define $Z^* = Z/W$ and $Y^* = Y \cup W$. What is the smallest set W such that $X \perp Z^*|Y^*$ is true? (2 points)

   Solution: $W = \{f\}$ is the smallest subset that satisfies the requirement. Y^* then is the Markov Blanket of node c.

9.3 Conditional Independence

From the graph, we can see that $a \perp c,d|b$. Prove using the axioms of probability that this implies $a \perp c|b$. (2 points)

Solution: $a \perp c,d|b$ means $P(a,c,d|b) = P(a|b)P(c,d|b)$. We want to prove $a \perp c|b$ using the axioms of probability.

$$
\begin{aligned}
P(a, c|b) &= \sum_d P(a, c, d|b) \text{ by the axiom of additivity for disjoint events} && (8)\\
&= \sum_d P(a|b)P(c, d|b) && (9)\\
&= P(a|b) \sum_d P(c, d|b) && (10)\\
&= P(a|b)P(c|b) && (11)
\end{aligned}
$$

Hence $a \perp c|b$. Other proofs are also accepted.

9.4 Structure learning

Suppose that we do not know the directionality of the edges $a - b$ and $b - c$, and we are trying to learn that by observing the conditional probability $p(a|b, c)$. Some of the entries in the table are observed and noted. Fill in the rest of the conditional probability table so that we obtain the directionality that we see in the graph, i.e, $a \to b$ and $b \to c$. (2 points)

P($a = 1|b = 0, c = 0$)	0.8
P($a = 1|b = 0, c = 1$)	**0.8**
P($a = 1|b = 1, c = 0$)	0.4
P($a = 1|b = 1, c = 1$)	**0.4**

Solution: We want $a \perp c|b$, i.e $P(a|b, c) = P(a|b)$. So we want that $P(a|b, c)$ should be the same for all values of c.