Homework 5 SOLUTIONS
CS 6375: Machine Learning

Spring 2012

Due date: Wednesday, May 2, 11:59 p.m.

1 Learning Theory [40 points, 10 points each]

- Mitchell 7.2

Solution:

a. Here, $|H| = \left(\frac{1}{2}(n+1)(n)\right)^2$, $n = 100$ and therefore

$$|H| = \left(\frac{1}{2}(100)(101)\right)^2$$

$\delta = 0.05$ and $\epsilon = 0.15$. Since the learner is consistent substituting the values of $|H|, \delta$ and ϵ in the following formula gives us the required value for m

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

b. The VC-dimension of axis-parallel rectangles in d dimensions is $2d$. Here $d = 2$ and therefore the VC dimension, $VC(H)$ is 4. Substituting $VC(H)$, $\delta = 0.05$ and $\epsilon = 0.15$ in the following expression gives us the required value for m:

$$m \geq \frac{1}{\epsilon}\left(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon)\right)$$

- Mitchell 7.3

Solution: Let us denote the true error by $error_{\mathcal{D}}$ and the training error on dataset D by $error_D$.

Chernoff bound:

$$\Pr(error_D < (1-\beta) error_{\mathcal{D}}(h)) \leq e^{-m \times error_{\mathcal{D}}(h) \beta^2/2}$$

Rearranging, we have:

$$\Pr(error_{\mathcal{D}}(h) > \frac{error_D(h)}{(1-\beta)}) \leq e^{-m \times error_{\mathcal{D}}(h) \beta^2/2} \quad (1)$$

We want an expression for

$$\Pr(error_{\mathcal{D}}(h) > error_D(h)(1+\gamma))$$

Comparing the two equations given above, we have:

$$\frac{1}{1-\beta} = 1+\gamma$$

which yields,

$$\beta = \frac{\gamma}{1+\gamma}$$

Note that to use the inequality, we must have $0 \leq \beta \leq 1$. This constrains the values that γ can take.

Substituting the value of β in Equation 1 we have,

$$\Pr(error_{\mathcal{D}}(h) > error_D(h)(1+\gamma)) \leq e^{-m \times error_{\mathcal{D}}(h) (\gamma/(1+\gamma))^2/2} \quad (2)$$

This gives us a bound on the probability of an arbitrarily chosen hypothesis.

To use it for computing sample complexity, we must consider the probability that any one of the $|H|$ hypothesis could have a large error

$$\Pr((\exists h \in H) error_{\mathcal{D}}(h) > error_D(h)(1+\gamma)) \leq |H| e^{-mp(\gamma/(1+\gamma))^2/2}$$

where $p = \min_{h \in H} error_{\mathcal{D}}(h)$.

If we call this probability δ, we get

$$\delta \leq |H| e^{-mp(\gamma/(1+\gamma))^2/2} \quad (3)$$

Taking log with base e on both sides, we get

$$\ln(\delta) \leq \ln(|H|) - mp(\gamma/(1+\gamma))^2/2 \tag{4}$$

Rearranging, we get,

$$m \geq \frac{2(1+\gamma)^2}{p\gamma^2}(\ln(|H|) + \ln(1/\delta))$$

- Mitchell 7.4

Solution: As we saw in class, the VC dimension of intervals is 2. This is because no set of 3 points can be shattered by an interval. Any set of 2 points can be shattered by intervals.

Substituting this dimension in the equation

$$m \geq \frac{1}{\epsilon}\left(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon)\right)$$

we get

$$m \geq \frac{1}{\epsilon}\left(4\log_2(2/\delta) + 16\log_2(13/\epsilon)\right)$$

Rearranging, we have:

$$\delta \geq \left[2^{4-m\epsilon}\left(\frac{13}{\epsilon}\right)^{16}\right]^{1/4}$$

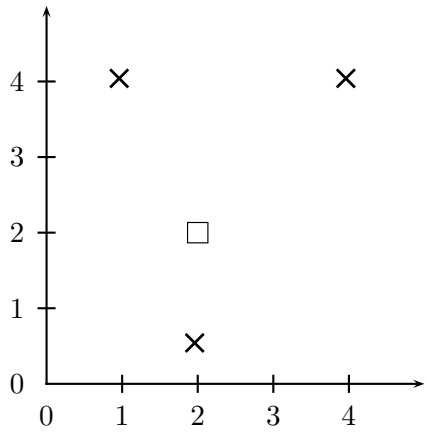
- Mitchell 7.5

Solution:

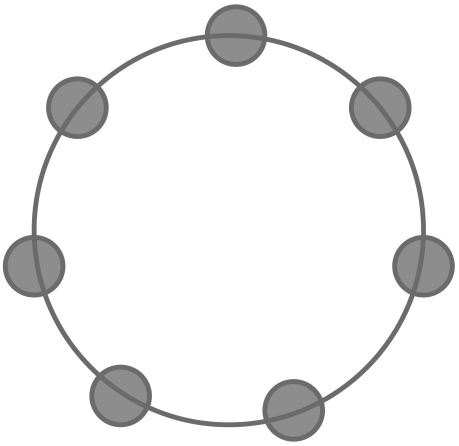
a. Consider the 6 points (1,0,0),(0,1,0),(0,0,1),(1,0,0),(0,1,0),(0,0,1). If we draw a bounding box for these points, then by excluding/including each point by moving a face of the box, we can get any labeling for the points. So the VC dimension is at least 6. For 7 points, consider the bounding box. If the bounding rectangle has at least one point in its interior, then we cannot accomplish the labeling where the interior point is labeled negative and the rest are labeled positive. If

none of the points are in the interior, then at least two must be on the same face of the rectangle by the piegon-hole principle and then cannot have opposite labels.

b. It is easy to see the VC-dimension of a circle is at least 3 since any 3 points on a triangle can be shattered. To prove that the VC dimension is not 4, we have to prove that no 4 points can be shattered and the formal proof is a little involved. Therefore, we will only show that some 4 points cannot be shattered. For instance, can you find a circle which covers all the crosses below but not the square.



c. The VC-dimension of a triangle is at least 7. All possible labelings of the seven points aligned on a circle can be separated using the triangles. See the figure below.



But we cannot find a placement of eight points whose all possible labelings are separable and this is quite hard to prove (requires knowledge of advanced geometry). Full credit if you prove that VC-dimension is ≥ 7.

- (**30 points**) Consider the following Bayesian network: $A \rightarrow B \rightarrow C$. And the following data table, with entries '?1' and '?2' missing at random:

A	B	C
F	F	F
F	F	?1
F	T	F
T	T	T
T	?2	T
T	F	T

  - Use the data to estimate initial parameters for this network, using maximum likelihood estimation for simplicity.
  - Apply the EM algorithm (by hand) to estimate the values of the missing data, reestimate the parameters, etc. until convergence. Show your calculations.
  - How many iterations does EM take to converge? Will this always be the case? Explain.

Solution:

1. The network has five parameters $P(A)$, $P(B|A)$, $P(B|\neg A)$, $P(C|B)$ and $P(C|\neg B)$. The ML estimates of these parameters are: $P(A) = 1/2$, $P(B|A) = 1/3$, $P(B|\neg A) = 1/3$, $P(C|B) = 1/2$ and $P(C|\neg B) = 1/3$.

2. We will start with the above parameters and run the EM algorithm.

   Iteration 1:

   - E-step: Complete the data using the current parameters. The second and the fifth example have two possible completions. There probabilities are given below:

     Completions of Example 2:

A	B	C	Prob
F	F	T	$\propto 1/2 \times 2/3 \times 2/3 = 2/3$
F	F	F	$\propto 1/2 \times 2/3 \times 1/3 = 1/3$

     Completions of Example 5:

A	B	C	Prob
T	T	T	0.428
T	F	T	0.571

   - M-step: Now the data set is bigger and weighted. The second and fifth examples are replaced by two weighted examples each given above. The weight associated with the remaining examples is 1. The MLE estimates based on the bigger, weighted dataset are: $P(A) = 0.5$, $P(B|A) = 0.476$, $P(B|\neg A) = 1/3$, $P(C|B) = 0.588$ and $P(C|\neg B) = 0.533$.

   Iteration 2:

   - E-step: Complete the data using the current parameters. The second and the fifth example have two possible completions. There probabilities are given below:

     Completions of Example 2:

A	B	C	Prob
F	F	T	0.533
F	F	F	0.466

     Completions of Example 5:

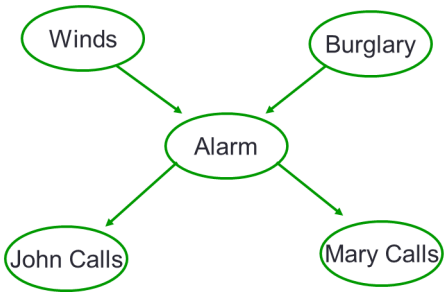
A	B	C	Prob
T	T	T	0.5
T	F	T	0.5

   - M-step: Now the data set is bigger and weighted. The second and fifth examples are replaced by two weighted examples each given above. The weight associated with the remaining examples is 1. The MLE estimates based on the bigger, weighted dataset are: $P(A) = 0.5$, $P(B|A) = 0.5$, $P(B|\neg A) = 1/3$, $P(C|B) = 0.6$ and $P(C|\neg B) = 0.5808$.

and so on. EM converges in about five iterations.

The parameters $P(A)$, $P(B|A)$, $P(B|\neg A)$, $P(C|B)$ and $P(C|\neg B)$ at convergence are 0.5, 0.5, 0.333333, 0.6, and 0.6 respectively.

3. EM will not always converge in five iterations. However, it will always converge to a local minima in finite number of steps. The number of steps required for convergence depends upon the problem and the initialization method.



- (**10 points**) Consider the Bayesian network given above. It has five variables: {Windy (W), Burglary(B), Alarm(A), John Calls(J), Mary Calls (M) }.
  - Is J independent of M?

    Solution: No.

  - Is B independent of W given A?

    Solution: No. The graph obtained by applying the dsep test is $W - A - B$. W and B are not disconnected in the graph.

  - Is M independent of W given A?

    Solution: Yes. The graph obtained by applying the dsep test is $W - A - B$ M. M and W are disconnected in the graph.

  - Is A independent of B given W?

Solution: No. The graph obtained by applying the dsep test is $W\ A - B$. A and B are not disconnected in the graph.

– Is B independent of J given A?

Solution: Yes. The graph obtained by applying the dsep test is $W - A - B$ J. J and B are disconnected in the graph.

2 Hidden Markov models [20 points]

Consider an HMM with three states, three outputs, and the following transition $P(X_{t+1}|X_t)$ and sensor $P(E_t|X_t)$ models. Assume a uniform distribution for the initial state, X_0.

X_t	X_{t+1}	a	b	c
a		0.5	0.4	0.1
b		0.1	0.5	0.4
c		0.4	0.1	0.5

X_t	E_t	p	q	r
a		0.7	0.1	0.2
b		0.2	0.7	0.1
c		0.1	0.2	0.7

1. Compute the most likely sequence of hidden states for the observed sequence, (p, p, r, r, q, r) by stepping through the Viterbi algorithm by hand. Show your work.

Solution:

p	p		r		r		q		r	
	0.116667		**0.040833**		**0.004083**		0.000408		0.000040	
0.233333	0.006667	0.081667	0.001867	0.008167	0.000327	0.000817	0.000033	0.000080	0.000023	0.000016
	0.013333		0.001067		0.002287		**0.000800**		**0.000080**	
	0.093333		**0.032667**		**0.003267**		**0.000327**		0.000032	
0.066667	0.033333	0.018667	0.009333	0.003267	0.001633	0.000327	0.000163	0.000229	**0.000114**	0.000011
	0.003333		0.000267		0.000572		0.000200		0.000020	
	0.023333		**0.008167**		0.000817		0.000082		0.000008	
0.033333	**0.026667**	0.002667	0.007467	0.005717	0.001307	0.002001	0.000131	0.000200	0.000091	0.000070
	0.016667		0.001333		**0.002858**		**0.001000**		**0.000100**	
1	2		3		4		5		6	

➔ a→a→c→c→c→c

2. Use the forward-backward algorithm to compute the probability distribution over states at position 3. Show your work.

Solution:

Same as the Viterbi except that we use sum-out instead of max-out and propagate forward upto slice 3 and backward upto 3.