CM146: Introduction to Machine Learning **Winter 2018**

Final Exam

Mar. 22^{nd}, 2018

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam booklet contains **five** problems.
- You have 150 minutes to earn a total of 100 points.
- **Besides giving the correct answer, being concise and clear is very important. To get the full credit, you must show your work and explain your answers.**

Good Luck!

Name and ID: (2 Point)

Name		/2
True/False Questions		/14
Hidden Markov Models		/9
Naive Bayes		/12
Kernels and SVM		/25
Short Answer Questions		/38
Total		/100

1 True or False [14 pts]

Choose either True or False for each of the following statements. For the statement you believe it is *False*, please give your brief explanation of it. Two points for each question. *Note: the credit can **only** be granted if your explanation for the false statement is correct. Also note, a negated statement is not counted as a correct explanation.*

(a) Training a k-class classification model using one-against-all is always faster than using one-vs-one because one-vs-one requires to train more binary classifiers.

Solution: False. Although one-vs-one requires to train k(k-1)/2 classifiers, the classifiers are trained on subsets of data. The time complexity comparison between these two approaches depends on the underlying training algorithm and data distribution.

(b) We would expect the support vectors to remain the same in general as we move from a linear kernel to higher order polynomial kernels.

Solution: False. There are no guarantees that the support vectors remain the same. For example, the feature vectors corresponding to polynomial kernels are non-linear functions of the original input vectors and thus the support points for maximum margin separation in the feature space can be quite different.

(c) In a mistake-driven algorithm such as the Perception algorithm, if we make a mistake on example x_i with label y_i, we update the weights w so that we can guarantee that we now predict y_i correctly.

Solution: False. There is no guarantee.

(d) Consider a classification problem with n features. The VC dimension of the corresponding (linear) SVM hypothesis space is larger than that of the corresponding logistic regression hypothesis space.

Solution: False. Both are linear model they have same VC dimension.

(e) A 3-layer neural network with non-linear activation functions can learn non-linear decision boundaries.

Solution: True.

(f) In AdaBoost, the weight associated with each weak learner can be negative (less than 0).

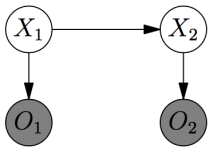
Solution: False. The weight describe the distribution of the data, and the probability cannot be negative.

(g) Using MAP to estimate model parameters always give us better performance.

Solution: False. If the prior is incorrect, the model performs worse.

2 Hidden Markov Models [9 pts]

Consider the following Hidden Markov Model.



X_1	$\Pr(X_1)$
0	0.3
1	0.7

X_t	X_{t+1}	$\Pr(X_{t+1}\|X_t)$
0	0	0.4
0	1	0.6
1	0	0.8
1	1	0.2

X_t	O_t	$\Pr(O_t\|X_t)$
0	A	0.9
0	B	0.1
1	A	0.5
1	B	0.5

Suppose that $O_1 = A$ and $O_2 = B$ is observed.

(a) **(3 pts)** What is the probability of $P(O_1 = A, O_2 = B, X_1 = 0, X_2 = 1)$?

Solution: 0.3*0.9*0.6*0.5 =0.081

(b) **(3 pts)** What is the most likely assignment for X_1 and X_2?

Solution: The most likely assignment given $O_1 = A, O_2 = B$ is $X_1 = 0, X_2 = 1$ You can use Viterbi, or list all four possibilities.

(c) **(3 pts)** [True/False] Based on the independent assumptions in HMM, the random variable O_1 is independent of the random variable X_2. Justify your answer.

Solution: False. Conditional independent does not imply independent.

3 Naive Bayes [12 pts]

Data the android is about to play in a concert on the Enterprise and he wants to use a Naive Bayes classifier to predict whether he will impress Captain Picard. He believes that the outcome depends on whether Picard has been reading Shakespeare or not for the three days before the concert. For the previous five concerts, Data has observed Picard and noted on which days he read Shakespeare. His observations look like this:

D1 (Day 1)	D2 (Day 2)	D3 (Day 3)	LC (LikedConcert)
1	1	0	yes
0	0	1	no
1	1	1	yes
1	0	1	no
0	0	0	no

(a) **(3 pts)** What does the modeling assumption make in the Naive Bayes model?

Solution: Given the label (LC) the days are independent, hence the joint probability can be written as $P(LC, D_1, D_2, D_3) = P(LC) \prod_1^3 P(D_i|LC)$

(b) **(4 pts)** Show the Naive Bayes model that Data obtains using maximum likelihood from these instances. (Write down the numerical values of the model parameters.)

Solution: LC is the top node, with arrows going to D1, D2 and D3. From the data, the max. likelihood probabilities are obtained by counting:
$P(LC = yes) = 2/5$
$P(D1 = 1 \mid LC = yes) = 1$
$P(D1 = 1 \mid LC = no) = 1/3$
$P(D2 = 1 \mid LC = yes) = 1$
$P(D2 = 1 \mid LC = no) = 0$
$P(D3 = 1 \mid LC = yes) = 1/2$
$P(D3 = 1 \mid LC = no) = 2/3$

(c) **(2 pts)** If Picard reads Shakespeare only on day 1 and day 2, how likely is he to enjoy Data's concert?

Solution: We need to compute

$$\begin{aligned} & P(LC = yes|D_1 = 1, D_2 = 1, D_3 = 0) \\ =& \frac{P(LC = 1, D_1 = 1, D_2 = 1, D_3 = 0)}{P(LC = yes, D_1 = 1, D_2 = 1, D_3 = 0) + P(LC = no, D_1 = 1, D_2 = 1, D_3 = 0)} \quad (1) \\ =& \frac{2/5 * 1 * 1 * 1/2}{2/5 * 1 * 1 * 1/2 + 3/5 * 0 * 1/3} = 1 \end{aligned}$$

(d) **(3 pts)** Estimate $P(LC = yes|D_2 = 1)$.

Solution: $P(LC = yes|D_2 = 1) = \frac{P(D_2=1|LC=yes) \cdot P(LC=yes)}{P(D_2=1)} = \frac{1 \cdot \frac{2}{5}}{\frac{2}{5}} = 1$

4 Kernels and SVM [25 pts]

(a) **(8 pts)** Properties of Kernels

i. **(3 pts)** Given n training examples $\{x_i\}_{i=1}^n$, the kernel matrix $\mathbf{A}$ is an $n \times n$ square matrix, where $\mathbf{A}(i,j) = K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$. Prove that the kernel matrix is symmetric (i.e, $A_{i,j} = A_{j,i}$).
hints: Your proof will not be longer than 2 or 3 lines.
Solution: We have $A_{ij} = K(x_1, x_2) = \Phi(x_1)^T \Phi(x_2) = \Phi(x_2)^T \Phi(x_1) = K(x_2, x_1) = A_{ji}$

ii. **(5 pts)** Prove that the kernel matrix $\mathbf{A}$ is positive semi-definite.
hints: (1) Remember that an $n \times n$ matrix $\mathbf{A}$ is positive semi-definite if and only if for any n dimensional vector $\mathbf{v} \neq \mathbf{0}$, we have $\mathbf{v}^T \mathbf{A} \mathbf{v} \geq 0$. (2) Consider a matrix $\mathbf{B} = [\Phi(x_1), \cdots, \Phi(x_n)]$ and use it to prove A is positive semi-definite.
Solution: Let $\Phi(x_i)$ be the feature map for the i^{th} example and define the matrix $\mathbf{B} = [\Phi(x_1), \cdots, \Phi(x_n)]$. It is easy to verify that $\mathbf{A} = \mathbf{B}^T \mathbf{B}$. Then, we have $\mathbf{v}^T \mathbf{A} \mathbf{v} = (\mathbf{B}\mathbf{v})^T \mathbf{B}\mathbf{v} = ||\mathbf{B}\mathbf{v}||^2 \geq 0$

(b) **(17 pts)** Given a dataset $D = \{x_i, y_i\}, x_i \in \mathbb{R}^k, y_i = \{-1, +1\}, 1 \leq i \leq N$.
A hard SVM solves the following formulation

$$\min_{w,b} \quad \frac{1}{2} w^T w \qquad \text{s.t} \quad \forall i, y_i(w^T x_i + b) \geq 1, \tag{2}$$

and soft SVM solves

$$\min_{w,\xi_i,b} \quad \frac{1}{2} w^T w + C \sum_i \xi_i \qquad \text{s.t} \quad \forall i, y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i, \xi_i \geq 0 \tag{3}$$

i. **(4 pts)** Complete:
If $C =$ _____, soft SVM will behave exactly as hard SVM.
In order to reduce over-fitting, one should _____ (decrease or increase) the value of C.
Solution: ∞, Decrease

ii. **(4 pts)** Show that when $C = 0$, the soft SVM returns a trivial solution and cannot be a good classification model.
Solution: When $C = 0$, ξ_i can be arbitrary large; therefore, the model ignores the constraints, and $w = 0$ is the optimal solution.

iii. **(3 pts)** [True/False] The slack variable ξ_i in soft SVM for a data point x_i always takes the value 0 if the data point is correctly classified by the hyper-plane. Explain your answer.
Solution: False. When the data point is correctly classified but inside the margin, ξ_i is non-zero.

iv. **(3 pts)** [True/False] The optimal weight vector w can be calculated as a linear combination of the training data points. Explain your answer. [You do not to prove this.]
Solution: True. Based on the dual representation.

v. **(3 pts)** We are given the dataset in Figure 1 below, where the positive examples are represented as black circles and negative points as white squares. (The same data is also provided in Table 1 for your convenience). Recall that the equation of the separating hyperplane is $\hat{y} = \mathbf{w}^T \mathbf{x} + b$.

i. Write down the parameters for the learned linear decision function.

$W = (w_1, w_2) =$ _____. $b =$ _____

Solution: $W = (1, 1)$, $b = -1$

ii. Circle all support vectors in Figure 1.

Solution: Point 1,7,8

index	x_1	x_2	y
1	0	0	−
2	0	-4	−
3	-1	-1	−
4	-2	-2	−
5	3	0	+
6	0	3	+
7	1	1	+
8	3	-1	+

Table 1: The dataset S

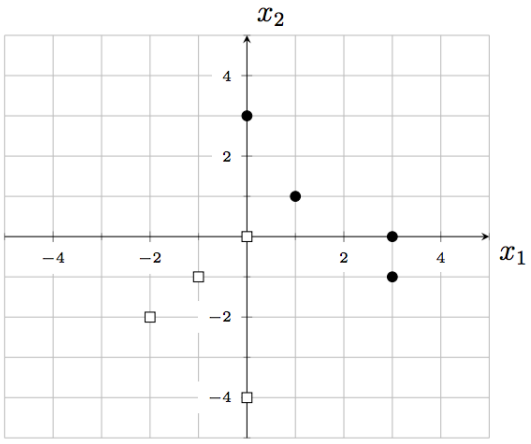


Figure 1: Linear SVM

5 Short Answer Questions [38 pts]

Most of the following questions can be answered in one or two sentences. Please make your answer concise and to the point.

(a) **(3 pts)** Consider training a classifier using AdaBoost with decision stumps (pick a horizontal or a vertical line, and one side of the half-space is positive and the other one is negative) on the following dataset:

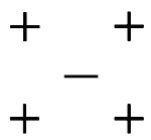


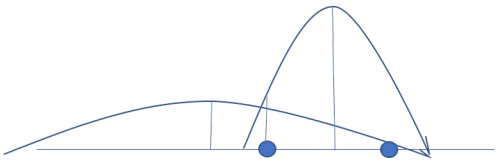
Figure 2: Example 2D dataset for Boosting

Which example(s) will have their weights increased at the end of the first iteration? Circle them and justify.

Solution: The negative example since the decision stump with least error in first iteration is constant over the whole domain. Notice this decision stump only predicts incorrectly on the negative example, whereas any other decision stump predicts incorrectly on at least two training examples.

(b) **(4 pts)** Suppose we clustered a set of N data points using two different clustering algorithms: k-means and Gaussian mixtures. In both cases we obtained 2 clusters and both algorithms return the same set of cluster centers. Can 2 points that are assigned to different clusters in the kmeans solution be assigned to the same cluster in the Gaussian mixture solution? If no, explain. If so, sketch an example and explain in 1-2 sentences.

Solution: Yes, it is possible. We can construct a case illustrate in the figure. One point is closer to the cluster center on the left. However, the distribution of the cluster on the right has lower variance, and it assigns higher probability on the point. Therefore, GMM group it with the cluster on the right while k-means assigns it to the left cluster.



(c) **(3 pts)** Suppose you are running a learning experiment on a new algorithm for binary classification. You have a data set consisting of 100 positive and 100 negative examples. You plan to use leave-one-out cross-validation (i.e., 200-fold cross-validation) to evaluate a baseline method: a simple majority function (i.e., returns the most frequent label on the training set as the prediction). What is the average cross-validation accuracy of the baseline? (Only need to write down the number).

Solution: 0.0

(d) **(4 pts)** P(Good Movie | Includes Tom Cruise) = 0.01
P(Good Movie | Tom Cruise absent) = 0.1
P(Tom Cruise in a randomly chosen movie) = 0.01

What is P(Tom Cruise is in the movie | Not a Good Movie)?

Solution: $T \sim$ Tom Cruise is in the movie, $G \sim$ Good Movie
$P(T|\tilde{G}) = \frac{P(T,\tilde{G})}{P(\tilde{G})} = \frac{P(\tilde{G}|T)P(T)}{P(\tilde{G}|T)P(T)+P(\tilde{G}|\tilde{T})P(\tilde{T})} = \frac{1}{91}$ Please do not send regrade requests if the denominator is 0.89. Even though in decimal the answer might be approximately the same, the denominator should not be 0.89

(e) **(6 pts)** We can easily extend the binary Logistic Regression model to handle multi-class classification. Lets assume we have K different classes, and posterior probability for class k is given as

$$P(y = k|X = x) = \frac{\exp(w_k^T x)}{\sum_{k'=1}^{K-1} \exp(w_{k'}^T x)} \tag{4}$$

where x is a d dimensional vector and w_k is the weight matrix for the k^{th} class.
Assuming dataset D consists of n examples, derive the log likelihood condition for this classifier.

hints: Let I_{ik} be an indicator function, where $i = 1, \ldots, n$ and $I_{ik} = \begin{cases} 1 & \text{if } y_i = k \\ 0 & \text{if } y_i \neq k \end{cases}$

(Full points if the derivation is mathematically correct. 2 points if you can describe the procedure for deriving.)

Solution:

$$L(w_1, \ldots, w_k) = \prod_{i=1}^{n} \prod_{k=1}^{K} P(y_i = k|X = x_i; w)^{I_{ik}}$$

$$= \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{\exp(w_k^T x_i)}{\sum_r \exp(w_r^T x_i)} \right)^{I_{ik}}$$

Taking logarithm,

$$l(w_1, \ldots, w_k) = \sum_{i=1}^{n} \sum_{k=1}^{K} I_{ik} \left[w_k^T x_i - \ln \sum_r \exp(w_r^T x_i) \right]$$

(f) **(6 pts)** In class we learned the following PAC learning bound for consistent learners:
Theorem 1. Let H be a finite concept class. Let D be an arbitrary, fixed unknown distribution over X. For any ϵ, $\delta > 0$, if we draw a sample S from D of size

$$m > \frac{1}{\epsilon}\left(ln(|H|) + ln\frac{1}{\delta}\right) \tag{5}$$

then with probability at least $1 - \delta$, all hypothesis $h \in H$ have $err_D(h) \leq \epsilon$. Our friend Kai is trying to solve a learning problem that fits in the assumptions above.

i. Kai tried a training set of 100 examples and observed some test error, so he wanted to reduce the test error to half. How many examples should Kai use, according to the above PAC bound?

Solution: 200, since we want to reduce ϵ to half.

ii. Kai took your suggestion and ran his algorithm again, however the error on the test set did not halve. Do you think it is possible? explain briefly.

Solution: Yes, it is possible. The theorem only shows the upper bound.

(g) **(4 pts)** List two differences between generative and discriminative learning models.

Solution: Generative models learn the joint probability $P(x, y)$, while discriminative model learns the conditional probability $P(Y|X)$. Generative models use Bayes rule in the inference, descriminative model directly compare the conditional probability. Generative model can be directly extended to the unsupervised setting by maximizing the marginal likelihood like in the EM algorithm. (There are many other differences.)

(h) **(8 pts)** We define a set of functions $T = f(x) = I[x > a] : a \in \mathbb{R}^1$, where $I[x > a]$ is the indicator function returning 1 if $x > a$ and returning 0 otherwise. For input domain $X = \mathbb{R}^1$, and a fixed positive number k, consider a concept class DT_k consisting of all decision trees of depth at most k where the function at each non-leaf node is an element of T. Note that if the tree has only one decision node (the root) and two leaves, then $k = 1$.

Determine the VC dimension of DT_3, and prove that your answer is correct.

Solution: First, note that the root node of the tree partitions the input space into two intervals. Each child node then recursively divides the corresponding interval into two more intervals. Hence, a full decision tree of depth k divides the input space into at most 2^k intervals, each of which can be assigned a label of 0 or 1. We will show that the VC dimension of DT_3 is 8.

Proof that $VC(DT_3) \geq 8$: given points, construct a decision tree such that each point lies in a separate interval. Then, one can assign labels to the leaves of the tree corresponding to any possible labeling of the points.

Proof that $VC(DT_k) < 2^3 + 1$: recall that a function from DT_3 can divide the input space into at most 8 intervals. Consequentially, the pigeonhole principle tells us that, given 9 points, at least one interval must contain two points no matter which function we are considering. This implies that no function in DT_3 can shatter 9 points, since every function in the class will contain an interval with more than one point and thus cannot assign different labels to those points.