10-701 Introduction to Machine Learning
Midterm Exam **Solutions**

Instructors: Eric Xing, Ziv Bar-Joseph

17 November, 2015

There are 11 questions, for a total of 100 points.
This exam is open book, open notes, but no computers or other electronic devices.
This exam is challenging, but don't worry because we will grade on a curve. Work efficiently.
Good luck!

Name: _____

Andrew ID: _____

Question	Points	Score
Basic Probability and MLE	10	
Decision Trees	10	
Naïve Bayes & Logistic Regression	6	
Deep Neural Networks	10	
SVM	12	
Bias-Variance Decomposition	14	
Gaussian Mixture Model	6	
Semi-Supervised learning	12	
Learning Theory, PAC learning	10	
Bayes Network	10	
Total	100	

1 Basic Probability and MLE - 10 points

1. You are trapped in a dark cave with three indistinguishable exits on the walls. One of the exits takes you 3 hours to travel and takes you outside. One of the other exits takes 1 hour to travel and the other takes 2 hours, but both drop you back in the original cave. You have no way of marking which exits you have attempted. What is the expected time it takes for you to get outside?

   Answer: Let the random variable, X be the time it takes for you to get outside. So, by the description of the problem, $\mathbb{E}(X) = \frac{1}{3}(3) + \frac{1}{3}(1 + \mathbb{E}(X)) + \frac{1}{3}(2 + \mathbb{E}(X))$. Solving this equation leads to the solution, $\mathbb{E}(X) = 6$.

2. Let $X_1, \cdots, X_n$ be iid data from a uniform distribution over the disc of radius θ in $\mathbb{R}^2$. Thus, $X_i \in \mathbb{R}^2$ and

$$p(x;\theta) = \begin{cases} \frac{1}{\pi\theta^2} & \text{if } \|x\| \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

   where $\|x\| = \sqrt{x_1^2 + x_2^2}$. Plese find the maximum likelihood estimate of θ.

   Answer: The likelihood function is

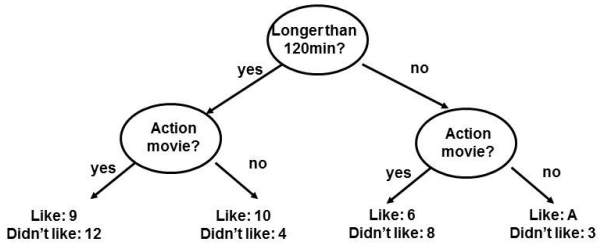
$$L(X^n;\theta) = \frac{1}{(\pi\theta^2)^n} \mathbf{1}\left\{ \max_{1\leq i\leq n} \|X_i\| \leq \theta \right\}$$

   When θ increases, the likelihood decreases. So θ should be as small as possible. However, we also need that $\theta \geq \max_{1\leq i\leq n} \|X_i\|$, otherwise the likelihood will drop to 0. So the maximum likelihood estimator is

$$\hat{\theta} = \max_{1\leq i\leq n} \|X_i\|$$

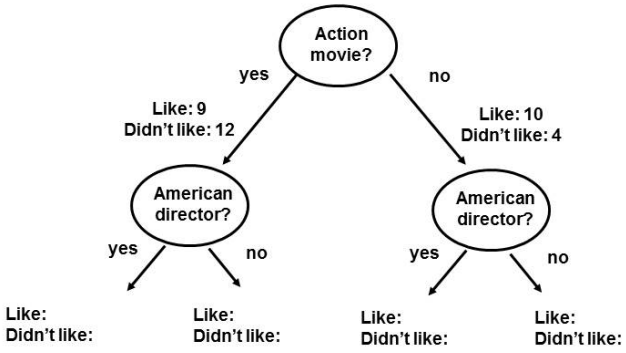
2 Decision Trees - 10 points

1. The following figure presents the top two levels of a decision tree learned to predict the attractiveness of a book. What should be the value of A if the decision tree was learned using the algorithm discussed in class (you can either say 'At most X' or 'At least X' or 'Equal to X' where you should replace X with a number based on your calculation), explain your answer?



Answer: A has to be 0. Action divides the inputs as: No (10+A,7), Yes(15,20) and Length would divide it as Yes(19,16), No(6+A,11). Any number higher than 0 would lead to a higher IG for Action and so it would be at the root.

2. We now focus on all samples assigned to the left side of the tree (i.e. those that are longer than 120 minutes). We know that we have a binary feature, 'American director' that after the 'Action movie' split provides a perfect split for the data (i.e. all samples on one side are 'like' and all those on the other side 'didn't like'. Fill in the missing values in the picture below:



Left: Yes Like: 9, Dont like: 0, No Like 0, Dont like 12
Right: Yes Like: 0, Dont like: 4, No Like 10, Dont like 0

3 Naïve Bayes & Logistic Regression - 6 points

1. In online learning, we can update the decision boundary of a classifier based on new data without reprocessing the old data. Now for a new data point that is an outlier, which of the following classifiers are likely to be effected more severely? NB, LR, SVM? Please give a one sentence explanation to your answer.

   Answer: NB and LR are likely to be affected more severely (answering either one is ok), because both are based on density estimation (or conditional probability estimation) that will be fit on the entire data, and every datapoint will carry a weight. On the other hand, SVM only depends on support vectors and controls the model complexity through a regularization term, which makes it more roboust to noise.

2. Now to build a classifier on discrete features using small training data, one will need to consider the scenario where some features have rare values that were never observed in the training data (e.g., the word 'Buenos Aires' does not appear in training for a text classification problem). To train a generalizable classifier, do you want to use NB or LR, how will you augment the original formulation of the classifier under a Bayesian or regularization setting?

   Answer: I would use NB based on word frequency, and augment the frequency with pseudo-count, which corresponds to a Dirichlet prior over the multinomial frequency in a Bayesian setting.

3. Now to build a classifier on high-dimensional features using small training data, one will need to consider the scenario where many features are just irrelevant noises. To train a generalizable classifier, do you want to use NB or LR, how will you augment the original formulation of the classifier under a Bayesian or regularization setting?

   Answer: I would use LR, and use an L1 regularization as it encourages sparisty in the coefficients of the model.

4 Deep Neural Networks - 10 points

In homework 3, we counted the model parameters of a convolutional neural network (CNN), which gives us a sense how much memory a CNN will consume. Now we estimate the computation overhead of CNNs by counting the FLOPs (floating point operations). For simplicity we only consider the forward pass.

Consider a convolutional layer C followed by a max pooling layer P. The input of layer C has 50 channels, each of which is of size 12×12. Layer C has 20 filters, each of which is of size 4×4. The convolution padding is 1 and the stride is 2. Layer P performs max pooling over each of the C's output feature maps, with 3×3 local receptive fields, and stride 1.

Given $x_1, x_2, \cdots, x_n$ all scalars, we assume:

- A scalar multiplication $x_i \cdot x_j$ accounts for one FLOP;
- A scalar addition $x_i + x_j$ accounts for one FLOP;
- A max operation $\max\{x_1, x_2, \cdots, x_n\}$ accounts for $n - 1$ FLOPs.

All other operations (e.g., $x_1 = x_2$) do not account for FLOPs.

Quetions:

1. What is size of each of layer P's output feature map?

   $4 \times 4 (\times 20)$

   (Layer C has 20 feature maps each of which is of size 6×6)

2. How many FLOPs layer C and P conduct in total during one forward pass?
   If account for bias: $20 \times (6 \times 6) \times (4 \times 4 \times 50 \times 2) + (8 \times 16 \times 20)$;
   otherwise: $20 \times (6 \times 6) \times (4 \times 4 \times 50 \times 2 - 1) + (8 \times 16 \times 20)$
   Either answer is OK.

5 SVM - 12 points

Recall that the soft-margin primal SVM problem is

$$\begin{array}{ll} \min & \frac{1}{2}w^T w + C\sum_{i=1}^{n}\xi_i \\ s.t. & \xi_i \geq 0, \forall i \in \{1, \cdots, n\} \\ & (w^T x_i + b)y_i \geq 1 - \xi_i, \forall i \in \{1, \cdots, n\} \end{array} \tag{1}$$

We can get the kernel SVM by taking the dual of the primal problem and then replace the product of $x_i^T x_j$ by $k(x_i, x_j)$ where $k(\cdot, \cdot)$ is the kernel function.

Figure 1 plots SVM decision boundaries resulting from using different kernels and/or different slack penalties. In Figure 1, there are two classes of training data, with labels $y_i \in \{-1, 1\}$, represented by circles and squares respectively. The SOLID circles and squares represent the support vectors. Label each plot in Figure 1 with the letter of the optimization problem below. You are NOT required to explain the reasons.

a) A soft-margin linear SVM with $C = 0.1$.

b) A soft-margin linear SVM with $C = 10$.

c) A hard-margin kernel SVM with $K(u, v) = u^T v + (u^T v)^2$.

d) A hard-margin kernel SVM with $K(u, v) = \exp(-\frac{1}{4}\|u - v\|_2^2)$.

e) A hard-margin kernel SVM with $K(u, v) = \exp(-4\|u - v\|_2^2)$.

f) None of the above.

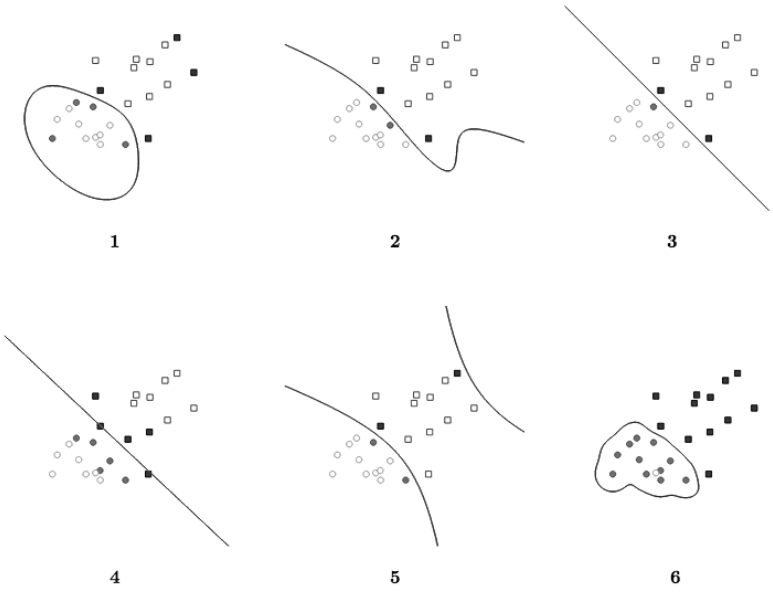


Figure 1: Induced Decision Boundaries

Answer: (a): 4. (b): 3. (c): 5. (d): 1. (e): 6. (f): 2.

Reason: (a). The decision boundary of linear SVM is linear. In comparison with Fig. 3 (corresponds to (b)), the line does not separate the two classes strictly, which corresponds to the case C is small and more errors are allowed.

(b). The decision boundary of linear SVM is linear. In com- parison with 4(corresponds to (a)), the line separates two classes strictly, which corresponds to the case C is big.

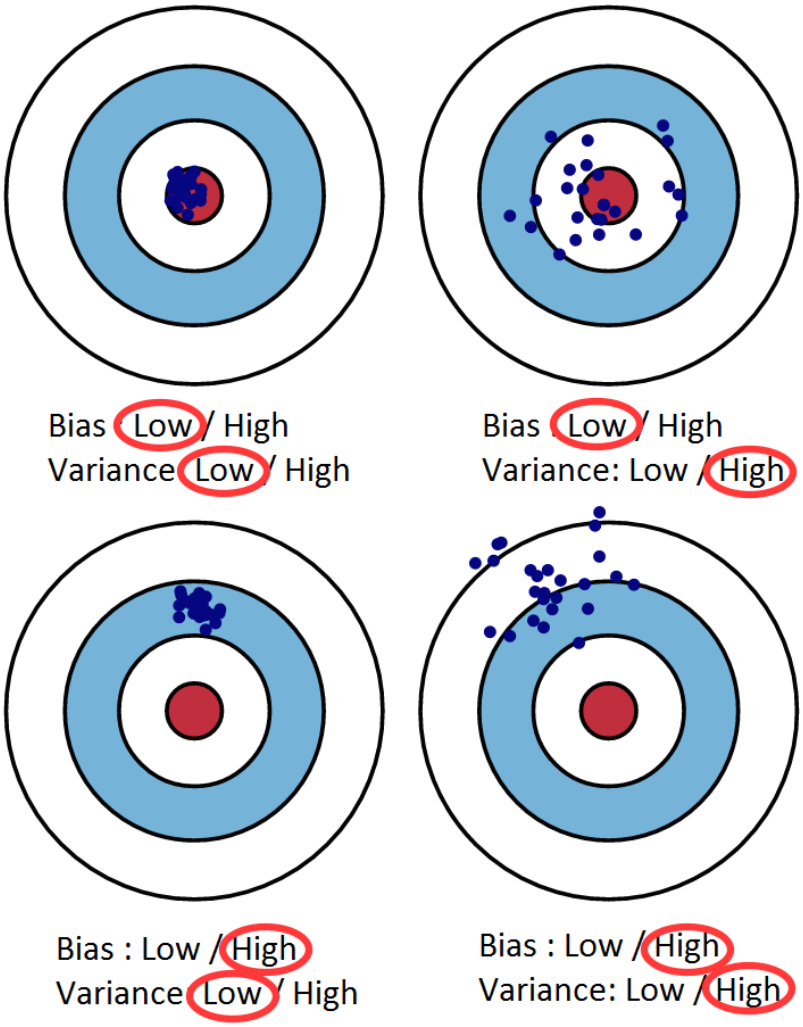
(c). The decision function of quadratic kernel is given by $f(x) = \sum_i \alpha_i (x_i \cdot x + (x_i \cdot x)^2) + b$. Hence the decision boundary is $f(x) = 0$. Since $f(x)$ is second order function of x, the curve can be ellipse or hyperbolic curve. Fig. 5 is hyperbolic curve.

(d). We can write out the decision function as $f(x) = \sum_i \alpha_i \exp(-\gamma_i \|x_i - x\|^2) + b$. If γ is large, then the kernel value is quite small even if the distance between x and x_i is small. This makes the classication hard with few supporting vectors. If Fig. 1 corresponds to the case γ is large (=4), then it is hard to classify many circle point in the middle in Fig. 1. Hence, Fig. 1 corresponds to $\gamma = \frac{1}{4}$.

(e). Using similar argument, we can conclude that if γ is large, there are more support vectors.

6 Bias-Variance Decomposition - 14 points

1. To understand bias and variance, we will create a graphical visualization using a bulls-eye. Imagine that the center of the target is our true model (a model that perfectly predicts the correct values). As we move away from the bulls-eye, our predictions get worse and worse. Imagine we can repeat our entire model building process to get a number of separate hits on the target. Each hit represents an individual realization of our model, given the chance variability in the training data we gather. Sometimes we will get a good distribution of training data so we predict very well and we are close to the bulls-eye, while sometimes our training data might be full of outliers or non-standard values resulting in poorer predictions. Consider these four different realizations resulting from a scatter of hits on the target. Characterize the bias and variance of the estimates of the following models on the data with respect to the true model as low or high by circling the appropriate entries below each diagram.



2. Explain what effect will the following operations have on the bias and variance of your model. Fill in one of 'increases', 'decreases' or 'no change' in each of the cells:

	Bias	Variance
Regularizing the weights in a linear/logistic regression model	increases	decreases
Increasing k in k-nearest neighbor models	increases	decreases
Pruning a decision tree (to a certain depth for example)	increases	decreases
Increasing the number of hidden units in an artificial neural network	decreases	increases
Using dropout to train a deep neural network	increases	decreases
Removing all the non-support vectors in SVM	no change	no change

7 Gaussian Mixture Model - 6 points

Consider a mixture distribution given by

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|z_k). \tag{2}$$

Suppose that we partition the vector $\boldsymbol{x}$ into two parts as $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, then the conditional distribution $p(\boldsymbol{x}_2|\boldsymbol{x}_1)$ is also a mixture distribution:

$$p(\boldsymbol{x}_2|\boldsymbol{x}_1) = \sum_{k=1}^{K} \epsilon_k p(\boldsymbol{x}_2|\boldsymbol{x}_1, z_k). \tag{3}$$

Give the expression of ϵ_k.

$$p(x_2|x_1) = \sum_k p(x_2, z_k|x_1) = \sum_k p(x_2|z_k, x_1)p(z_k|x_1) = \sum_k p(x_2|z_k, x_1)p(x_1|z_k)p(z_k)/p(x_1)$$

$$\epsilon = \pi_k p(x_1|z_k)/p(x_1)$$

8 Semi-Supervised learning - 12 points

1. We would like to use semi-supervised learning to classify text documents. We are using the 'bag of words' representation discussed in class with binary indicators for the presence of 10000 words in each document (so each document is represented by a binary vector of length 10000).

   For the following classifiers and learning methods discussed in class, state whether the method can be applied to improve the classifier (Yes) or not (No) and provide a brief explanation.

   (a) (Yes / No) Naïve Bayes using EM
   Brief explanation:
   Yes. NB is a probabilistic method and so can use the EM approach

   (b) (Yes / No) Naïve Bayes using co-training
   Brief explanation:
   No. No way to split the data into two independent sets of features.

   (c) (Yes / No) Naïve Bayes using model complexity selection
   Brief explanation:
   No. We do not use feature transformation so impossible to use this approach for this classifier.

   (d) (Yes / No) Decision trees using re-weighting
   Brief explanation:
   Yes. Re-weighting will impact the IG for each feature.

   (e) (Yes / No) Decision tree using EM
   Brief explanation:
   No. Not a probabilistic method.

   (f) (Yes / No) Decision trees using model complexity selection
   Brief explanation:
   Yes. We can use to select tree depth, for example.

2. Unlike all other classifiers we discussed, KNN does not have any parameters to tune. For each of the following semi-supervised methods state whether a KNN classifier (where K is fixed and not allowed to change) learned for some data using labeled and unlabeled data could be different from a KNN classifier learned using only the labeled data in this dataset (no need to explain).

   (a) Reweighting: Same / Different
   Different. Will impact the identity of the neighbors since one of the neighbors can be weighted more than 1 and so others would be excluded.

   (b) Co-training: Same / Different
   Different. Would provide more labeled points.

   (c) Model complexity selection: Same / Different
   Same. The only parameter is k and its fixed so this method would not have an impact on the resulting classifier.

9 Learning Theory, PAC learning - 10 points

In class we learned the following agnostic PAC learning bound:

Theorem 1. *Let H be a finite concept class. Let D be an arbitrary, fixed unknown distribution over X. For any $\epsilon, \delta > 0$, if we draw a sample S from D of size*

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right), \tag{4}$$

then with probability at least $1-\delta$, all hypothesis $h \in H$ have $|\mathsf{err}_D(h) - \mathsf{err}_S(h)| \leq \epsilon$.

Our friend Yan is trying to solve a learning problem that fits in the assumptions above.

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

   400, since we want to reduce ϵ to half.

2. Yan took your suggestion and ran his algorithm again, however the overfitting did not halve. Do you think it is possible? Explain briefly.

   It is possible. Reasons include 1) sample complexity is a probabilistic statement rather than a deterministic one, and 2) test error is only an estimate of the true error. Stating either of them will count as correct.

10 Bayes Networks - 10 points

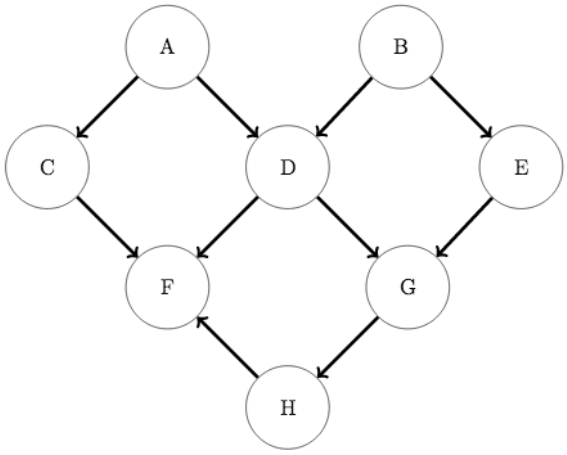


Figure 2: A graphical model

Consider the Bayesian network in Figure 2. We use $(X \perp\!\!\!\perp Y|Z)$ to denote the fact that X and Y are independent given Z. Answer the following questions:

1. Are there any pairs of point that are independent? If your answer is yes, please list out all such pairs.
   Answer: Yes. (C, E), (C, B), (A, E), (A, B).

2. Does $(B \perp\!\!\!\perp C|A, D)$ hold? Briefly explain.

   Answer: It holds. We can see that by using the Bayes ball algorithm.

3. Does $(B \perp\!\!\!\perp F|A, D)$ hold? Briefly explain.

   Answer: It does not hold. By using the Bayes ball algorithm, we can find a path $B \to E \to G \to H \to F$.

4. Assuming that there are $d = 10$ values that each of these variables can take (say 1 to 10), how many parameters do we need to model the full joint distribution without using the kowldge incoded in the graph (i.e. no independence / conditional independence assumptions)? How many parameters do we need for the Bayesian network for such setting? (you do not need to provide the exact number, a close approximation or a tight upper / lower bound will do).

   Answer: Without graph: 10^8. With graph: about $10 \times 2 + 10^2 \times 3 + 10^3 \times 2 + 10^4 \times 1 = 12320$