Introduction to Machine Learning 10-701
Midterm, Tues April 8

[1 point]

Name:

Andrew ID:

Instructions:

- You are allowed a (two-sided) sheet of notes.
- Exam ends at 2:45pm
- Take a deep breath and don't spend too long on any one question!

	Max	Score
Name & Andrew id	1	
True/False	40	
Short Ques	59	
Total	100	

1 True or False questions [40 pts = 2 × 20]

Answer True or False. Justify your answer very briefly in 1-2 sentences.

1. Gaussian Mixture Model is essentially a probability distribution.

   True. GMM specifies the following distribution $p(x) = \sum_{i=1}^{K} \pi_i \mathcal{N}(\mu_i, \Sigma_i)$.

2. When the feature space is larger, overfitting is less likely.

   False. The more the number of features, the higher the complexity of the model and hence greater its ability to overfit the training data.

3. The EM iterations never decrease the likelihood function of the data.

   True. Both the E and M steps maximize a lower bound on the likelihood function of the data, and hence never decrease it.

4. Non-parametric models do not have parameters.

   False. Non-parametric models can have parameters e.g. kernel regression has the bandwidth parameter, but the number of parameters scale with the size of the dataset.

5. In kernel density estimation, a large kernel bandwidth will result in low bias.

   False. A large kernel bandwidth results in more smoothing and poor approximation, resulting in higher bias.

6. Non-parametric models are usually more efficient than parametric models in terms of model storage.

   False. Non-parametric models either need to look at the entire dataset to predict the label of test points or require the number of parameters to scale with the dataset size, hence require more storage.

7. Increasing the regularization parameter λ in lasso regression leads to sparser regression coefficients.

   True. Larger regularization parameter penalizes non-zero coefficients more, leading to sparser solution.

8. Boosting decision stumps can result in a quadratic decision boundary.

   False. The sign of a finite linear combination of decision stumps always results in a piecewise linear decision boundary.

9. Decision trees are generative classifiers.

   False. Decision trees do not assume a model for the input feature distribution, hence are not generative.

10. The following statement always holds for any joint probability distribution:

    $$p(X_1, X_2, X_3, \ldots, X_N) = p(X_1)p(X_2|X_1)p(X_3|X_1, X_2), \ldots p(X_N|X_1, X_2, \ldots, X_{N-1})$$

    True. This simply follows from chain rule.

11. For a fixed size of the training and test set, increasing the complexity of the model always leads to reduction of the test error.

    False. Increasing complexity of model for a fixed training and test set leads to overfitting the training data and reduction in *training* error, but not test error.

12. Suppose you wish to predict age of a person from his/her brain scan using regression, but you only have 10 subjects and each subject is represented by the brain activity at 20,000 regions in the brain. You would prefer to use least squares regression instead of ridge regression.

    False. When the number of datapoints (subjects) is less than number of features, the least squares solution needs to be regularized to prevent overfitting, hence we prefer ridge regression.

13. The Local Markov Property states a Node is independent of non-descendants given at least one of its parents.

    False. The Local Markov Property states a Node is independent of non-descendants given *all* of its parents.

14. HMM is a generative model.

    True. HMM assumes a model for the data generating process.

15. The goal of the Viterbi algorithm is to learn the parameters of the Hidden Markov Model.

    False. Baum-Welch algorithm is used to learn the parameters of an HMM, Viterbi is used to infer the most likely state assignment.

16. When doing kernel regression on a memory-constrained device, you should prefer to use a box kernel instead of a Gaussian kernel.

    True. A box kernel only uses a few data points for prediction and hence does not need to load the entire dataset into memory unlike Gaussian kernel which assigns non-zero weight to all training data points.

17. To predict the chance that Steelers football team will win the Super Bowl Championship next year, you should prefer to use logistic regression instead of decision trees.

    True. Logistic regression will characterize the probability (chance) of label being win or loss, whereas decision tree will simply output the decision (win or loss).

18. The kmeans algorithm finds the global optimum of the kmeans cost function.

    False. The kmeans cost function is non-convex and the algorithm is only guaranteed to converge to a *local* optimum.

19. Let $X \in \mathbb{R}^{n \times m}$ be the data matrix of m n-dimensional data instances. Let $X = UDV$ denote the singular decomposition of X where D is a diagonal matrix containing the singular values ordered from largest to smallest. The 1st principal component of X is the first column vector of V.
    False. The 1st principal component of X is the first column vector of U.

20. The goal of independent component analysis is to transform the datapoints with a linear transformation in such a way that they become linearly independent.

    False. ICA transforms the datapoints with a linear transformation in such a way that they become *statistically* and not linearly independent.

2 Short questions [59 pts]

2.1 Loss functions [5pts]

Match each predictor to the loss function it typically minimizes:

boosting	$\Leftrightarrow$	exponential loss
logistic regression	$\Leftrightarrow$	log loss
k-NN classifier	$\Leftrightarrow$	0/1 loss
linear regression	$\Leftrightarrow$	squared loss
SVM classification	$\Leftrightarrow$	hinge loss

2.2 Decision boundaries [10 pts]

For each classifier, circle the type(s) of decision boundary it can yield for a binary classification problem. In some cases, more than 1 option may be correct. Circle all options that you think are correct.

decision trees:	linear, piecewise linear
(non-kernel) SVM:	linear
logistic regression	linear
Gaussian Naive Bayes	linear, quadratic
boosting	linear, piecewise linear, quadratic

2.3 Independence [6 pts]

Give an example of a 2D distribution where the marginal distributions are uncorrelated, but they are statistically dependent by construction.

There are many good solutions. Possible solutions: α-degree rotated 2D uniform distribution for $0 < \alpha < \pi/2$, or uniform distribution on a disk. Discrete solutions can also be accepted if it is proved that the marginal distributions are uncorrelated, but they are not independent.

2D Gaussian is not a good solution!

2.4 Leave-one-out error [6pts]

What is the leave-one-out error of

1. 1NN: 2 or (2/4)
2. 3 NN 1 or (1/4)

on the following dataset:

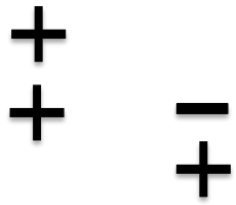
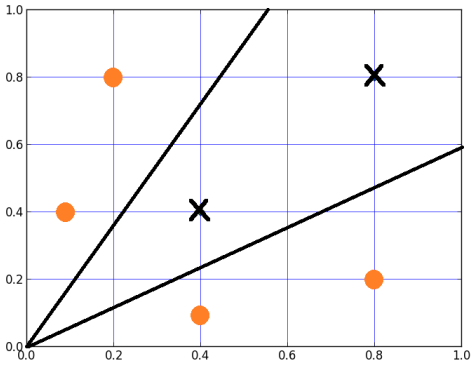


Figure 1: Figure for Leave-one-out Cross-validation

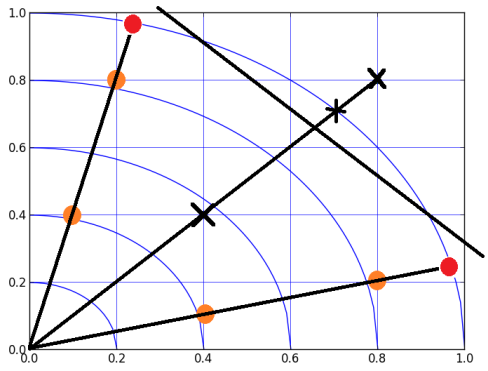
2.5 SVM kernels [10 pts]

The following figure (a) shows 6 labeled training data points from two classes.

(a)



(b)



1. Let's consider the following kernel $K(x,z) = \frac{z^T x}{\|x\| \, \|z\|}$. Prove that $K(x,z)$ is a kernel.
   Let $\phi(x) = \frac{x}{\|x\|}$. Then $K(x,z) = \phi(x)^T \phi(z)$, therefore K is a kernel function with

feature map ϕ.

2. Map each data point in the original feature space as shown in figure (a) to the new feature space representation implicit by $K(x, z)$ in figure (b). Drawn.

3. Is the training data linearly separable in the new feature space? [Yes? or No?] Yes.

4. In figure (b), draw the decision boundary for the maximum margin separator in this implicit feature space. A qualitative drawing is sufficient. Drawn.

5. In figure (a), draw the decision boundary in the original feature space resulting from this kernel. A qualitative drawing is sufficient. Drawn.

2.6 Graphical Model [8 pts]

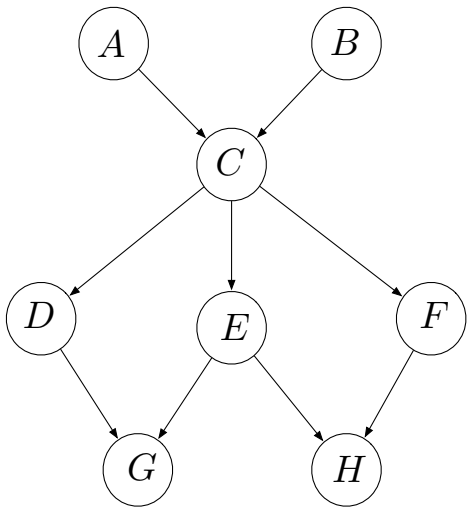


Figure 2: Bayesian network for Graphical Model question

For this question, refer to the graphical model in Figure 2.

- What is the factorization of the joint probability distribution $p(A, B, C, D, E, F, G, H)$ according to this graphical model.

$$p(A)p(B)p(C \mid A, B)p(D \mid C)p(E \mid C)p(F \mid C)p(G \mid D, E)p(H \mid E, F)$$

- Assume that every random variable is a discrete random variable with K possible values. How many parameters are needed to represent this graphical model?

$$\mathcal{O}(2K + 3K^3 + 3K^2)$$

- Is the following statement true or false? Explain why.
  - $A \perp H$

    False. The path $A \rightarrow C \rightarrow F \rightarrow H$ does not meet the D-separation criterion (other solutions are possible too).
  - $A \perp H|B, D, F$

    False. The path $A \rightarrow C \rightarrow E \rightarrow H$ does not meet the D-separation criterion when conditioned on the set B, D, F.

2.7 Hidden Markov Models [6 pts]

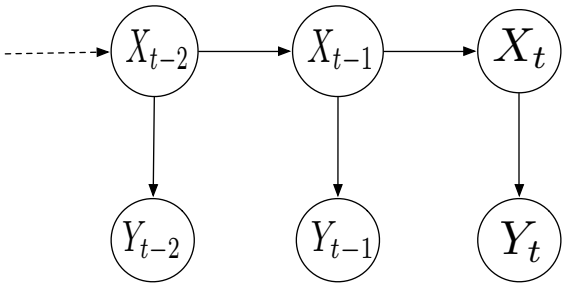


Figure 3: Figure for the HMM question.

Answer true or false. Explain your reasoning in 1 sentences.

- Refer to Figure 3. Suppose that we know the parameters of this HMM. Having observed X_{t-2} and X_{t-1}, we can make better prediction for X_t than observing only X_{t-1}.

  False. X_t is conditionally independent of X_{t-2} given X_{t-1}, therefore X_{t-2} cannot provide more information to predict X_t given X_{t-1}.

- Refer to Figure 3. Suppose that we don't know the parameters of this HMM. Having observed Y_{t-2} and Y_{t-1}, we can make better prediction for Y_t than observing only Y_{t-1}.

  True. More training data can help to learn better parameters which can lead to better prediction.

- Exact inference with variable elimination is intractable in (first-order) HMMs in general.

  False. It is a tree, so exact inference is tractable.

2.8 MLE [8 pts]

We have a random variable X drawn from a Poisson distribution. The Poisson distribution is a discrete distribution and X can be any positive integer. The probability of X at a point x is $p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$. Given points $x_1, \ldots, x_n$, write down the MLE estimate of λ.

$$
\begin{aligned}
p(x_1, \ldots, x_n|\lambda) &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\
\log p(x_1, \ldots, x_n|\lambda) &\propto (\sum_{i=1}^{n} x_i) \log \lambda - n\lambda \\
0 &= \frac{\partial \log p(x_1, \ldots, x_n|\lambda)}{\partial \lambda} = (\sum_{i=1}^{n} x_i) \frac{1}{\lambda} - n \\
\hat{\lambda} &= \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}
$$