CS 7641 CSE/ISYE 6740 Final Exam Solution (2013 Fall)

Le Song

12/10 Tue, 2:50 - 5:40 pm

- Name:
- GT ID:
- E-mail:

Problem	Point	Your Score	Problem	Point	Your Score
1	20		6	20	
2	20		7	30	
3	20		8	30	
4	20		9	20	
5	20		Total	200	

Instructions:

- Try your best to be clear as much as possible. No credit may be given to unreadable writing.
- The exam is open book and open note, but no electronic devices (including smart phones) are allowed.
- Please use the back-side of each sheet if you are out of space.
- Good luck!

1 A Big Picture [20 pts]

We learned the following machine learning methods during the semester. **Use the number (1 to 14)** to answer the questions below.

1. AdaBoost
2. Decision trees
3. Gaussian mixture
4. Histogram
5. K-nearest neighbor
6. K-means
7. Kernel density estimation
8. Linear regression
9. Logistic regression
10. Naive Bayes
11. Neural networks
12. Principal component analysis
13. Singular value decomposition
14. Support vector machines

(a) List all methods which are used for classification. [4 pts]

Answer: 1, 2, 5, 9, 10, 11, 14

(b) List all methods which are used for regression. [4 pts]

Answer: 2, 5, 8, 11, 14

(c) List all methods which are used for density estimation. [4 pts]

Answer: 4, 7

(d) List all methods which are used for clustering. [4 pts]

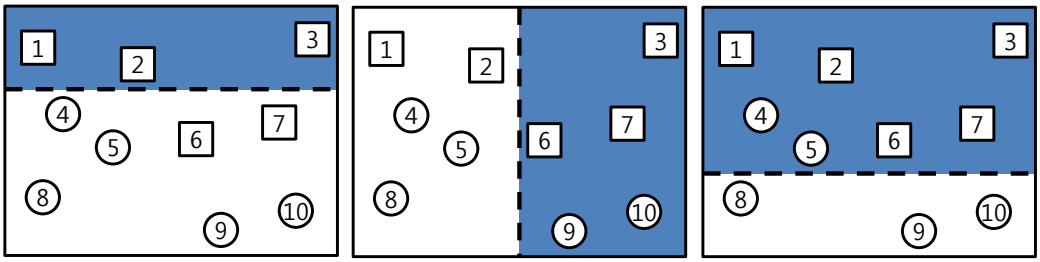
Answer: 3, 6

(e) List all methods which are used for dimension reduction. [4 pts]

Answer: 12, 13

2 Boosting [20 pts]

In this problem, we test your understanding of AdaBoost algorithm with a simple binary classification example. We are given 10 data points, belonging to either the `square` class or the `circle` class. The following figures show the decision boundary of three weak learners (h_1, h_2, and h_3). Suppose we boost with these weak learners in that order. In the figure, the darkened region means it is classified as `square` class, while white region indicates it is classified as `circle` class by the corresponding weak learner. We use the same notations with the lecture note: h_t is t-th weak learner, $\epsilon_t = \sum_{i=1}^{n} D_t(i) I\{y^i \neq h_t(x^i; \theta)\}$, $\alpha_t = 1/2 \ln(\frac{1-\epsilon_t}{\epsilon_t})$, and $D_t(i)$ is weight of data point i for t-th weak learner.



(a) When we learn the second weak learner h_2, list data points which receives higher weights than others. [4 pts]

Answer: 6, 7

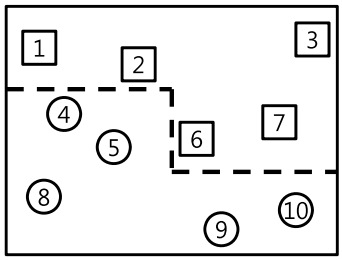
(b) When we learn the third weak learner h_3, list data points which receive smallest weights. [4 pts]

Answer: 3, 4, 5, 8

(c) What is ϵ_1? [4 pts]

Answer: 0.2

(d) Draw the final decision boundary in the figure below. [4 pts]



(e) What is the classification error with train data of the final classifier? [4 pts]

Answer: 0

3 Decision Trees [20 pts]

The following dataset will be used to learn a decision tree for predicting whether a mushroom is edible or not based on its shape, color and odor.

Shape	Color	Odor	Edible
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

(a) What is entropy $H(Edible|Odor = 1$ or $Odor = 3)$? Use log base 2. [5 pts]

Answer: 1

(b) Which attribute would the decision tree algorithm using multi-way split choose to use for the root of the tree, assuming we do not post-prune the tree? [5 pts]

Answer: Odor

(c) Draw the full decision tree that would be learned for this data (no pruning). [7 pts]

Answer: Odor(1:Y, 3:N, 2:Color(W:Y, G/U:N, B:Shape(C:Y, D:N)))

(d) Suppose we have a validation set as follows. What will be the training set error and validation set error of the tree? Express your answer as the number of examples that would be misclassified. [3 pts]

Shape	Color	Odor	Edible
C	B	2	No
D	B	2	No
C	W	2	Yes

- Train error: 0 / 11
- Validation error: 1 / 3

4 Kernels [20 pts]

This section asks your basic understanding about kernels. For each of the following kernels, prove or disprove whether it is a valid kernel. x and y are arbitrary real-valued vectors in d dimension Euclidean space.

(a) $K_1(x, y) = -1$ **[5 pts]**

- Answer: No
- Reason: Suppose $z = (z_1, z_2, ..., z_d)$. Then, the gram matrix $\mathcal{K}_1$ will be filled with -1 in every cell. $z^\top \mathcal{K}_1 z$ will be a $d \times d$ matrix filled with $-z_1^2 - z_2^2 - ... - z_d^2 < 0$. As the gram matrix is non-positive definite, this is not a kernel.

(b) $K_2(x, y) = (1 + x^\top y)^2$ **[5 pts]**

- Answer: Yes
- Reason:

$$
\begin{aligned}
K_2(x, y) &= (1 + x^\top y)^2 \\
&= 1 + 2x^\top y + (x^\top y)(x^\top y) \\
&= 1 + \sum_{i=1}^{d} \left(2x_i y_i + x_i^2 y_i^2\right) + 2 \sum_{j>k} x_j x_k y_j y_k \\
&= [1, \sqrt{2}x_1, ..., \sqrt{2}x_d, x_1^2, ..., x_d^2, x_1x_2, x_1x_3, ..., x_{d-1}x_d]^\top [1, \sqrt{2}y_1, ..., \sqrt{2}y_d, y_1^2, ..., y_d^2, y_1y_2, y_1y_3, ..., y_{d-1}y_d]
\end{aligned}
$$

With $\phi(x) = [1, \sqrt{2}x_1, ..., \sqrt{2}x_d, x_1^2, ..., x_d^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, ..., \sqrt{2}x_{d-1}x_d]$, we have $K_2(x, y) = \phi(x)^\top \phi(y)$.

(c) $K_3(x, y) = f(x)f(y)$**, where** f **is some real-valued function. [5 pts]**

- Answer: Yes
- Reason: First, it is obviously symmetric as $f(x)f(y) = f(y)f(x)$. Suppose $z \in \mathbb{R}^d$ and $\mathcal{K}_3$ is the gram matrix with this kernel. Then,

$$\begin{aligned} z^\top \mathcal{K}_3 z &= \sum_{i=1}^{d}\sum_{j=1}^{d} z_i z_j f(x^i) y(x^i) \\ &= \sum_{i=1}^{d} z_i^2 f(x^i)^2 + 2\sum_{j>k} z_j z_k y(x^j) y(x^k) \\ &= \left(\sum_{i=1}^{d} z_i f(x^i)\right)^2 \geq 0 \end{aligned}$$

  Thus, the gram matrix is positive semi-definite.

(d) $K_4(x, y) = K_5(x, y)K_6(x, y)$**, where** K_5 **and** K_6 **are some valid kernels. [5 pts]**

- Answer: Yes
- Reason: As both K_5 and K_6 are valid kernels, we can express each of them as an inner product, namely, $K_5(x, y) = \varphi(x)^\top \varphi(y)$ and $K_6(x, y) = \psi(x)^\top \psi(y)$. Thus,

$$\begin{aligned} K_4(x, y) &= (\varphi(x)^\top \varphi(y))(\psi(x)^\top \psi(y)) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} [\varphi(x)]_i [\varphi(y)]_i [\psi(x)]_j [\psi(y)]_j \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} ([\varphi(x)]_i [\psi(x)]_j)([\varphi(y)]_i [\psi(y)]_j) \end{aligned}$$

  where m and n are the dimension of $\varphi(x)$ and $\psi(y)$, respectively. With $\phi(x) = \{[\varphi(x)]_i [\psi(x)]_j\}$ for all possible combinations of $i = 1, ..., m$ and $j = 1, ..., n$, we have $K_4(x, y) = \phi(x)\phi(y)$.

5 Model Selection [20 pts]

We learned about bias-variance decomposition and model selection in the class. It basically deals with the problem of choosing a model complexity. In each sub-questions, we show a pair of two candidate models for various machine learning problems. Mark B on the one with less bias, and mark V on the other. You are not required to explain why. For example,

- Model 1: A model with less bias – (B)
- Model 2: A model with less variance – (V)

(a) [3 pts]

- Model 1: A flexible model with many parameters – (B)
- Model 2: A rigid model with a few parameters – (V)

(b) [3 pts]

- Model 1: Ridge regression with large regularization coefficient λ – (V)
- Model 2: Unregularized linear regression – (B)

(c) [3 pts]

- Model 1: Regression with higher degree – (B)
- Model 2: Regression with lower degree – (V)

(d) [3 pts] For a logistic regression $p(y = 1|x, \theta) = \frac{1}{1+\exp\{-\theta^\top x\}}$**,**

- Model 1: Logistic regression with large θ – (B)
- Model 2: Logistic regression with small θ – (V)

(e) [4 pts]

- Model 1: K-nearest Neighbor with large K – (V)
- Model 2: K-nearest Neighbor with small K – (B)

(f) [4 pts]

- Model 1: A classifier with irregular decision boundary – (B)
- Model 2: A classifier with smooth decision boundary – (V)

6 Maximum Likelihood [20 pts]

(a) Uniform distribution [10 pts]

A uniform distribution in the range of $[0, \theta)$ is given by

$$p(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x < \theta \\ 0 & \text{otherwise} \end{cases}.$$

What is the maximum likelihood estimator of θ?

Hint: Please be careful. This is not the same as the question in the mid-term.

Answer: There is no MLE for this distribution. Think of the largest observed data point $\max(x_i)$. As we can't observe a data at θ, $\hat{\theta}_{ML}$ can't be $\max(x_i)$. One may argue that $\hat{\theta}_{ML} + \delta$ for very small δ can be the MLE, but we can always find ϵ such that $0 < \epsilon < \delta$. That is, we can always find another $\hat{\theta}$ which leads to higher likelihood of the given data. Thus, we do not have MLE in this case.

(b) Dependent Noise Model [10 pts]

Let $X_1, ..., X_n$ be n determinations of a physical constant θ. Consider the model,

$$X_i = \theta + e_i, \quad i = 1, ..., n$$

and assume

$$e_i = \alpha e_{i-1} + \epsilon_i, \quad i = 1, ..., n, \quad e_0 = 0$$

with ϵ_i is i.i.d standard normal, and α is a known constant. What is the maximum likelihood estimate of θ?

Answer: According to the definition above, $X_1 = \theta + \epsilon_1$, and $X_i = \theta + \alpha e_{i-1} + \epsilon_i = \theta + \alpha(X_{i-1} - \theta) + \epsilon_i$ for $i > 1$. As e_i is i.i.d normal (with standard deviation σ^2), $p(X_1) \sim \mathcal{N}(\theta, \sigma^2)$ and $p(X_i) \sim \mathcal{N}((1-\alpha)\theta + \alpha X_{i-1})$ for $i > 1$.

Likelihood function of θ is given by $L(\theta) = p(X_1)p(X_2|X_1)p(X_3|X_2)...p(X_n|X_{n-1})$ and its log-likelihood is $\ell(\theta) = \log p(X_1) + \log p(X_2|X_1) + \log p(X_3|X_2) + ... + \log p(X_n|X_{n-1})$. Using the definition of normal distribution,

$$\ell(\theta) = \ln \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X_1 - \theta)^2}{2\sigma^2}\right\} + \sum_{i=2}^{n} \ln \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X_i - \alpha X_{i-1} - (1-\alpha)\theta)^2}{2\sigma^2}\right\}$$

Taking partial derivatives w.r.t θ and setting it to 0 gives

$$\frac{\partial \ell(\theta)}{\partial \theta} = (X_1 - \theta) + \sum_{i=2}^{n} (X_i - \alpha X_{i-1} - (1-\alpha)\theta) = 0$$

Solving it by θ gives

$$\hat{\theta} = \frac{(1 - \alpha + \alpha^2)X_1 + \sum_{i=2}^{n-1}(1-\alpha)^2 X_i + (\alpha - 1)X_n}{1 + (n-1)(1-\alpha)^2}$$

7 Naive Bayes and Bayes Classification [30 pts]

Consider a classification problem with the table of observations below. X_1 and X_2 are two binary random variables which are the observed variables. Y is the class label which is observed for the training data given below. We will use the Naive Bayes classifier and the Bayes classifier to classify a new instance after training on the data below and compare the results.

X_1	X_2	Y	Counts
0	0	0	2
0	0	1	18
1	0	0	4
1	0	1	1
0	1	0	4
0	1	1	1
1	1	0	2
1	1	1	18

(a) Construct the Naive Bayes classifier given the data above. Use it to classify the instance $(0, 0)$**. [5 pts]**

Answer:

$P(Y = 0|X_1 = 0, X_2 = 0) \propto P(Y = 0) \times P(X_1 = 0|Y = 0) \times P(X_2 = 0|Y = 0) = 0.24 \times 0.5 \times 0.5 = 0.06$
$P(Y = 1|X_1 = 0, X_2 = 0) \propto P(Y = 1) \times P(X_1 = 0|Y = 1) \times P(X_2 = 0|Y = 1) = 0.76 \times 0.5 \times 0.5 = 0.19$

So the instance will be classified as $y = 1$.

(b) Construct the Bayes classifier given the data above. Use it to classify the instance $(0, 0)$**. [5 pts]**

Answer:

$P(Y = 0|X_1 = 0, X_2 = 0) \propto P(Y = 0) \times P(X_1 = 0, X_2 = 0|Y = 0) = 0.24 \times 1/6 = 0.04$
$P(Y = 1|X_1 = 0, X_2 = 0) \propto P(Y = 1) \times P(X_1 = 0, X_2 = 0|Y = 1) = 0.76 \times 9/19 = 0.36$

So the instance will be classified as $y = 1$.

(c) Compare the number of independent parameters in the two classifiers. Instead of just two observed data variables, if there were n random binary observed variables $\{X_1, \ldots, X_n\}$, what would be the number of parameters required for both classifiers? From this, what can you say about the rate of growth of the number of parameters for both models as n increases? [10 pts]

Answer:

Number of independent parameters in the naive bays model is $1 + 4 = 5$.
Number of independent parameters in the bays model is $1 + 2 \times 3 = 7$.
If we have n observed variables, the umber of independent parameters in the naive bays model is $1 + 2n$. The number of independent parameters in the naive bays model is $1 + 2(2^n - 1)$.

(d) Compute the probabilities $P_{NB}(y = 1|x_1 = 0, x_2 = 0)$ using the Naive Bayes classifier and $P_{Bayes}(y = 1|x_1 = 0, x_2 = 0)$ using the Bayes classifier. Explain why $P_{NB} \neq P_{Bayes}$. [10 pts]

Answer:

$P_{NB}(y = 1|x_1 = 0, x_2 = 0) = 0.19/(2/5) = 0.475$
$P_{Bayes}(y = 1|x_1 = 0, x_2 = 0) = 0.36/(2/5) = 0.9$

They are different so the Naive Bayes assumption of class-conditional independence of features is violated.

8 Hidden Markov Model [30 pts]

Consider a Hidden Markov Model (HMM) with states $Y_t \in \{S_1, S_2, S_3\}$, observations $X_t \in \{A, B, C\}$, and parameters in the following table. Note that in state transition matrix, a_{ij} indicates the probability to move from state i to state j.

Initial state	State transition probability			Emission probability		
$\pi_1 = 1$	$a_{11} = 1/2$	$a_{12} = 1/4$	$a_{13} = 1/4$	$b_1(A) = 1/2$	$b_1(B) = 1/2$	$b_1(C) = 0$
$\pi_2 = 0$	$a_{21} = 0$	$a_{22} = 1/2$	$a_{23} = 1/2$	$b_2(A) = 1/2$	$b_2(B) = 0$	$b_2(C) = 1/2$
$\pi_3 = 0$	$a_{31} = 0$	$a_{32} = 0$	$a_{33} = 1$	$b_3(A) = 0$	$b_3(B) = 1/2$	$b_3(C) = 1/2$

(a) What is $P(Y_5 = S_3)$? [5 pts]

Answer: $P(Y_5 = S_3) = 1 - P(Y_5 = S_1) - P(Y_5 = S_2) = 1 - 1/16 - 4/32 = 13/16$

(b) What is $P(Y_5 = S_3 | X_{1:7} = AABCABC)$? [5 pts]

Answer: 0

(c) Fill in the following table assuming the observation $AABCABC$. The α's are values obtained during the forward algorithm : $\alpha_t(i) = P(X_1, \ldots, X_t, Y_t = i)$. [10 pts]

t	$\alpha_t(1)$	$\alpha_t(2)$	$\alpha_t(3)$
1	$\frac{1}{2}$	0	0
2	$\frac{1}{8}$	$\frac{1}{16}$	0
3	$\frac{1}{32}$	0	$\frac{1}{32}$
4	0	$\frac{1}{256}$	$\frac{5}{256}$
5	0	$\frac{1}{2^{10}}$	0
6	0	0	$\frac{1}{2^{12}}$
7	0	0	$\frac{1}{2^{13}}$

(d) Write down the sequence of $Y_{1:7}$ with the maximal posterior probability assuming the observation $AABCABC$. What is that posterior probability? [10 pts]

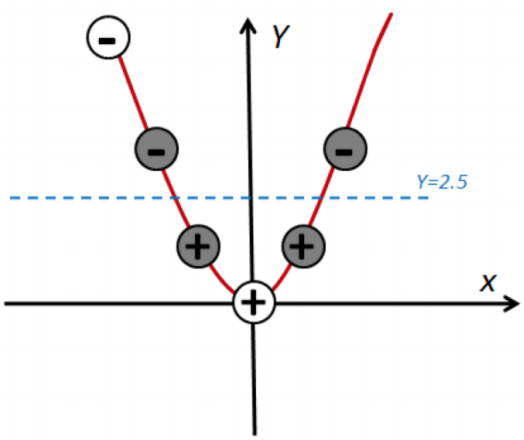
Answer: $S_1S_1S_1S_2S_2S_3S_3$ with posterior probability 1.

9 Support Vector Machines [20 pts]

Suppose we have a dataset in 1-d space which consists of 3 data points $\{-1, 0, 1\}$ with the positive label and 3 data points $\{-3, -2, 2\}$ with the negative label.

(a) Find a feature map ($\mathbb{R}^1 \to \mathbb{R}^2$), which will map the original 1-dimensional data points to the 2-dimensional feature space so that the positive samples and the negative samples are linearly separable with each other. Draw the dataset after mapping in the 2-dimensional space. [10 pts]

Answer: $x \to (x, x^2)$.



(b) In your plot above, draw the decision boundary given by hard-margin linear SVM. Mark the corresponding support vectors. [5 pts]

Answer: See the above figure.

(c) For the feature map you use, what is the corresponding kernel $K(x_1, x_2)$? [5 pts]

Answer: $K(x_1, x_2) = x_1 x_2 + (x_1 x_2)^2$.