CS188 Final Exam

Zhiyuan Yang

TOTAL POINTS

58 / 65

QUESTION 1

1 True/False (1 - 4) 8 / 8

- 0 Correct

QUESTION 2

2 True/False (5 - 8) 8 / 8

- 0 Correct

QUESTION 3

3 True/False (9 - 10) 2 / 4

- 2 problem 9 incorrect

QUESTION 4

4 Multiple Choice (11 - 15) 15 / 15

- 0 Correct

QUESTION 5

5 Short answers 16 1 / 2

- 1 part b incorrect

QUESTION 6

6 Short answers 17 3 / 4

- 1 Missing concrete solution (c)

QUESTION 7

7 Short answers 18 4 / 4

- 0 Correct

QUESTION 8

8 Short answers 19 3 / 4

- 1 (b) Should be 90/900

QUESTION 9

9 Short answers 20 2 / 4

- 1 a) Computation error
- 1 b) Reasoning for maximization vs. min

QUESTION 10

10 Short answers 21 4 / 4

- 0 Correct

QUESTION 11

11 Short answers 22 8 / 8

- 0 Correct

CS 188 — Introduction to Machine Learning: Final

Winter 2017

Name: Zhiyuan Yang

UID: 304 618 600

Instructions

1. This exam is CLOSED BOOK and CLOSED NOTES.
2. The time limit for the exam is **3 hours**.
3. Mark your answers ON THE EXAM ITSELF IN THE SPACE PROVIDED. If you make a mess, clearly indicate your final answer (box it).
4. DO NOT write on the reverse side.
5. You may use scratch paper if needed.
6. For true/false questions, CIRCLE True OR False
7. For multiple-choice questions, CIRCLE ALL CORRECT CHOICES AND ONLY THE CORRECT CHOICES (in some cases, there may be more than one but always at least one correct choice) for full credit.
8. For all other questions, show the work that you did to arrive at your answer so that we can give you partial credit where appropriate.
9. If you think something about a question is open to interpretation, feel free to ask the instructor or make a note on the exam.

Q	Problem	Points	Score
1-10	True/False	20	
11-15	Multiple choice	15	
16	Training and Validation	2	
17	Kernels	4	
18	Regularization	4	
19	Evaluation	4	
20	Generalizing linear regression	4	
21	Hidden Markov Models	4	
22	Clustering	8	
Total		65	

True/False

1. (2 pts) (True/False) In a single iteration of the Adaboost algorithm, the weights on all the misclassified points increase by the same multiplicative factor.

True

$\varepsilon_t = \sum W_t(n) \mathbb{1}(y_n \neq h_t(x_n))$

$\beta_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$ $\quad W_t(n) \propto W_t(n) e^{-\beta_t y_n h_t(x_n)}$

2. (2 pts) (True/False) In PCA, the first or top principal component corresponds to the direction of smallest variance.

False It corresponds to the direction of largest variance.

3. (2 pts) (True/False) Consider a mixture model of two 1-dimensional Gaussians, where the mixture distribution of $x \in \mathcal{R}$ is given by

True

$$P(x|\theta) = \sum_{j=1}^{2} \omega_j \mathcal{N}(x|\mu_j, \sigma_j^2),$$

Here $\mathcal{N}(x|\mu_j, \sigma_j^2)$ says that x is drawn from a Gaussian distribution with mean μ_j and variance σ_j^2. We can identify the most likely **posterior** assignment, i.e., j that maximizes $P(z = j|x)$ where z denotes the cluster membership of x by comparing the values of $\omega_1 \mathcal{N}(x; \mu_1, \sigma_1^2)$ and $\omega_2 \mathcal{N}(x; \mu_2, \sigma_2^2)$.

$P(z=j|x) = \frac{P(x|z=j) W_j}{\sum_{j'=1}^{2} P(x|z=j') W_{j'}}$

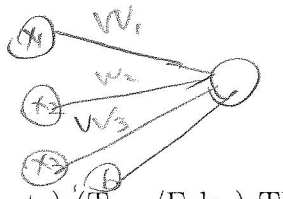
4. (2 pts) (True/False) Let $d(\boldsymbol{x}_n, \boldsymbol{x}_m)$ denote the Euclidean distance between vectors $\boldsymbol{x}_n$ and $\boldsymbol{x}_m$. Running K-Nearest Neighbors with distance measure $e^{d(\boldsymbol{x}_n, \boldsymbol{x}_m)}$ instead of $d(\boldsymbol{x}_n, \boldsymbol{x}_m)$

True $f(x) = e^x$ is a monotonically increasing function.

produces identical classification results.

5. (2 pts) (True/False) A neural network with no hidden layers and a single unit in the output layer with a sigmoid activation function is equivalent to a logistic regression model.

True



$\sigma(w^T x_n + b)$

↑ exactly logistic regression

6. (2 pts) (True/False) The EM algorithm is guaranteed to converge to a global maximum of the log likelihood.

False

It's guaranteed to converge to a local extrema.

7. (2 pts) (True/False) After normalizing the features (subtract each feature by its mean and divide by its standard deviation), the predictions of the learning algorithm will no longer depend on the units used to measure the features.

True

Normalization does exactly that

8. (2 pts) (True/False) To choose the hyperparameter C in SVMs, we learn a SVM for different values of C on a training set. Then, we evaluate each SVM (obtained for a fixed C) on a test set and choose the C whose model has the best accuracy on the test set. The accuracy of the chosen SVM on the test set is a good estimate of its generalization

False We need to do cross-validation on a separate validation set to optimize the parameter C.

accuracy.

9. (2 pts) (True/False) For a convex function f, if we find an input $\boldsymbol{x}_0$ such that $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$, then $\boldsymbol{x}_0$ is a global minimum of f.

False We also need that $H_f(X)$ is positive-definite.

10. (2 pts) (True/False) For a linear hypothesis $h_{\boldsymbol{w},b}(x) = \boldsymbol{w}^T\boldsymbol{x} + b$, the distance of the origin from the decision boundary is the same for hypotheses with parameters $(\boldsymbol{w}, b)$ and $(c\boldsymbol{w}, cb)$ for any $c > 0$.

True. Mathematically (w, b) and (cw, cb) define the same hyperplane/decision boundary.

Multiple choice

b

11. (3 pts) We are given data for 10 instances each of which has 5 features. The eigenvalues of the covariance matrix are $(20, 5, 0, 0, 0)$. What is the fraction of variance retained if we use only the first principal component to represent each individual?

    (a) 0
    (b) $\frac{20}{25}$
    (c) $\frac{5}{25}$
    (d) 1

12. (3 pts) Which of the following properties must hold for a kernel matrix ?

    (a) Symmetric
    (b) Invertible
    (c) All entries are non-negative
    (d) All eigenvalues are non-negative.

$K = \begin{pmatrix} \phi(x_1)^T\phi(x_1) & \cdots \\ & \phi(x_m)^T\phi(x_m) \end{pmatrix}$

13. (3 pts) For the same training data, you learn (unregularized) linear regression as well as ridge regression for a fixed hyperparameter $\lambda > 0$. Which of the following statements is true about the optimal value of the residual sum of squares (RSS) cost function for either model on the training data?

    (a) The RSS of ridge regression is never lower than the RSS for linear regression.
    (b) The RSS of linear regression is never lower than the RSS for ridge regression.
    (c) There are some datasets on which the RSS of linear regression is lower than ridge regression and others with ridge regression having lower RSS.
    (d) The RSS of both linear and ridge regression is the same on all datasets.

14. (3 pts) In learning which of these models is it important to consider solutions obtained from multiple random initializations?

    (a) A deep neural network with multiple hidden layers all of which have nonlinear activation functions
    (b) Logistic regression

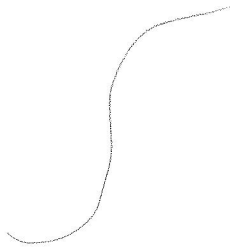
↑ randomly initialize weights and biases

$\min \frac{1}{2}\|w\|^2$ $\max L(x, \alpha, \beta)$

(c) Soft-margin SVM
(d) Gaussian Mixture Models → EM

15. (3 pts) Which of the following statements is true of the sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$?

(a) approaches 1 as x becomes large and positive
(b) approaches -1 as x becomes large and negative
(c) takes a value of $\frac{1}{2}$ at $x = 0$
(d) increases with increasing x



Short answers

16. (2 pts) **Training and Validation**

Figure 1 depicts the training and validation curves of a learner with increasing model complexity.

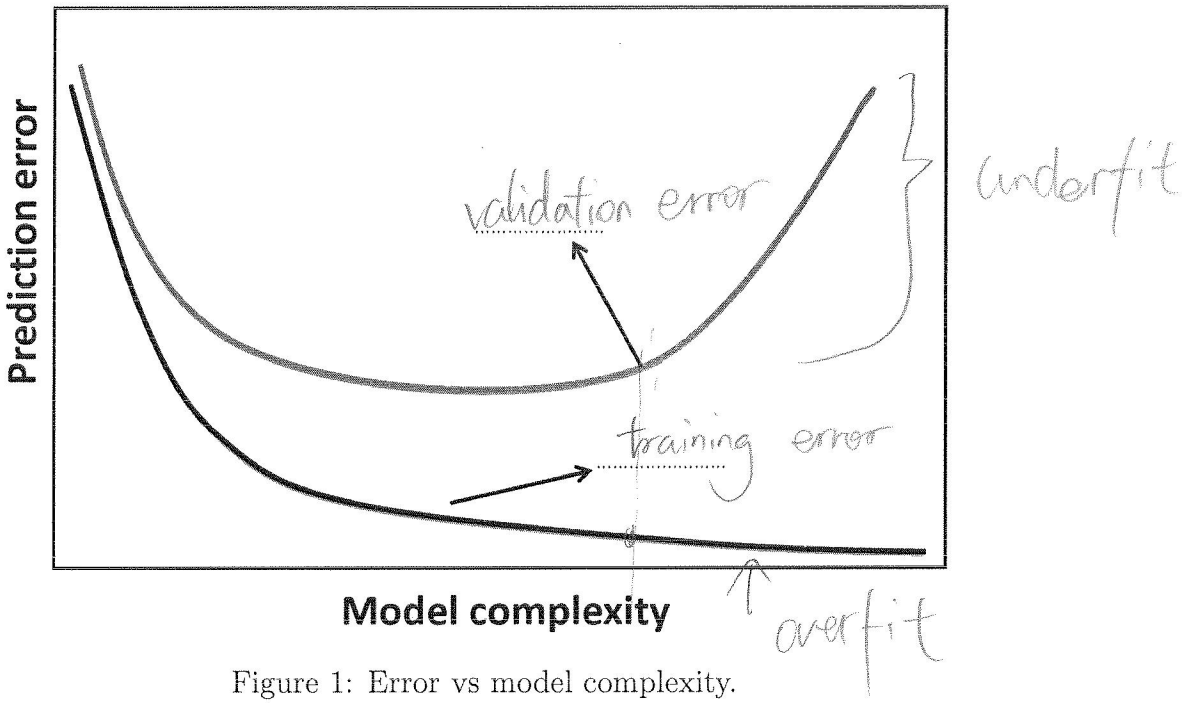


Figure 1: Error vs model complexity.

(a) (1 pts) Which of the curves is more likely to be the training error and which is more likely to be the validation error? Indicate on the figure by filling in the dotted lines.

(b) (1 pts) In which regions does the model overfit or underfit? Indicate clearly on the graph by labeling "overfit" and "underfit".

17. (4 pts) **Kernels**
You experiment with the following kernels in a soft-margin SVM framework where each of the input vectors $\boldsymbol{x} \in \mathbb{R}^2$:

(1) $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{x}'$
(2) $K(\boldsymbol{x}, \boldsymbol{x}') = 1 - 3(\boldsymbol{x}^T\boldsymbol{x}')^3$
(3) $K(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T\boldsymbol{x}' + 2)^5$
(4) $K(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T\boldsymbol{x}') + 2(\boldsymbol{x}^T\boldsymbol{x}')^2$

(a) (1 pts) What is the primary motivation for the use of kernels ?

So we can generalize our model to non-linear spaces and merely need to compute dot products between feature vectors without computing the actual mapping

(b) (1 pts) The optimization routine for the dual SVM problem complained that one of the kernels is not valid. Which one (1-4)? 2

(c) (2 pts) Prove that your choice in (b) is indeed not a valid kernel.

To be a valid kernel, K needs to be positive-definite. That is,

$$z^T K z > 0 \quad \forall z$$

$$z^T \left(1 - 3(x^T x')^3\right) z$$

$$= z^T z - 3 z^T (x^T x')^3 z$$

$$= \left(1 - 3(x^T x')^3\right) \|z\|^2$$

↑ no guarantee.

18. (4 pts) **Regularization**
The cost function for L_2-regularized linear regression is

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta},$$

where $\lambda > 0$.

(a) (2 pts) Suppose we accidentally write

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda\boldsymbol{y}^T\boldsymbol{y}$$

instead. Explain why this form of "regularization" has no effect.

If we take the gradient of the wrong cost function w.r.t. θ, $\lambda y^T y$ goes away so won't contribute to our estimating optimal $\hat{\theta}$

(b) (2 pts) Suppose we use the correct expression but accidentally choose $\lambda < 0$. Explain briefly how this defeats the purpose of regularization.

We apply regularization because we prefer simpler models than complex ones — we want θ to have small values. If we choose $\lambda < 0$, $\theta \to \infty$ would minimize the cost function, which defeats our purpose completely

19. (4 pts) **Evaluation**
You run an algorithm to classify spam versus non-spam emails. Your test data contains 900 non-spam (negative instances) and 100 spam (positive instances) emails. The classifier predicts 90 true positives and 90 false positives.

(a) (1 pts) What is the true positive rate of the classifier (you can leave your answer in fractions) ?

Actual \ Pred.	1	0
1	90	10
0	90	810

$$TPR = \frac{90}{100}$$

(b) (1 pts) What is the false positive rate of the classifier (you can leave your answer in fractions)?

$$FPR = \frac{10}{900}$$

(c) (1 pts) What fraction of the emails predicted as spam by the classifier are truly spam (you can leave your answers in fractions) ?

$$\frac{90}{180}$$

(d) (1 pts) What is the term used to describe the measure of performance that you computed in part (c)?

Precision

20. (4 pts) **Generalizing linear regression: Laplace regression**
In class, we showed how linear regression (ordinary least squares) can be interpreted as a probabilistic model. In this problem, we will explore how probabilistic models allow us to generalize learning algorithms. In each of these examples, we have a training set $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ of N independent instances where $x_n \in \mathbb{R}^D$.

Now we model our target y_n as coming from adding Laplace noise to a hyperplane. Specifically

$$p\left(y_n|\boldsymbol{x}_n;\boldsymbol{\theta}\right) = \frac{1}{2b}\exp\left(-\frac{|y_n - \boldsymbol{\theta}^T\boldsymbol{x}_n|}{b}\right)$$

(a) (2 pts) Write the log likelihood of the parameters $l(\boldsymbol{\theta})$. Express your answer in terms of y_n, $\boldsymbol{x}_n$, $\boldsymbol{\theta}$, and b.

$$L = (2b)^{-N}\exp\left(-\frac{1}{b}\Sigma|y_n - \theta^T x_n|\right)$$

$$LL = -N\log(2b) + \log\left(-\frac{1}{b}\Sigma|y_n - \theta^T x_n|\right)$$

$$= -N\log 2b - \log b + \log\Sigma|y_n - \theta^T x_n|$$

$$\left(= \text{constant} + \log\Sigma|y_n - \theta^T x_n|\right)$$

(b) (2 pts) Show that the maximum likelihood estimate of $\boldsymbol{\theta}$ is the $\boldsymbol{\theta}$ that minimizes a new cost function: the *sum of absolute residuals*:

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} |y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n|$$

From part (a),

$$\hat{\theta}_{mle} = \text{argmin } LL(\theta) = \text{argmin } \left(\text{constant} + \sum |y_n - \theta^T x_n|\right)$$

$$= \text{argmin } \sum |y_n - \theta^T x_n|$$

$$= \text{argmin } J(\theta)$$

21. (4 pts) **Hidden Markov Models**

You are keen on monitoring your health over the summer. Your health state reflects if you have a cold (state = 2) or you are feeling well (state = 1). Inspired by CS 188, you seek to model your health with an HMM (because your health state on a given day is correlated with your health state the next day). There are two possible output symbols, $L = low$ or $H = high$ energy.

For this, you need to specify the model parameters $\theta = \{\boldsymbol{\pi}, \boldsymbol{Q}, \boldsymbol{E}\}$, where $\boldsymbol{\pi}$ is the initial state distribution, $\boldsymbol{Q}$ are the state transition probabilities, and $\boldsymbol{E}$ are the emission probabilities. You determine the parameters of the HMM as:

$$\boldsymbol{\pi} = \begin{matrix} 1 \\ 2 \end{matrix}\begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \quad \boldsymbol{Q} = \begin{matrix} \\ 1 \\ 2 \end{matrix}\begin{matrix} \begin{matrix} 1 & 2 \end{matrix} \\ \begin{bmatrix} 1-a & a \\ a & 1-a \end{bmatrix} \end{matrix} \quad E = \begin{matrix} \\ 1 \\ 2 \end{matrix}\begin{matrix} \begin{matrix} L & H \end{matrix} \\ \begin{bmatrix} a & 1-a \\ 1 & 0 \end{bmatrix} \end{matrix}$$

where, e.g., $\pi_1 = P(X_1 = 1)$, $q_{21} = P(X_{t+1} = 2 | X_t = 1)$, and
$e_1(L) = P(Y_t = L | X_t = 1) = a$,
$e_1(H) = P(Y_t = H | X_t = 1) = 1 - a$,
$e_2(L) = P(Y_t = L | X_t = 2) = 1$,
$e_2(H) = P(Y_t = H | X_t = 2) = 0$.

We also have $0 < a < 1$.

(a) (2 pts) What is the probability that you are well on the first day (state = 1) and then get a cold (state = 2) the next day. That is, what is $P(X_1 = 1, X_2 = 2)$ (Your final answer must involve numbers and the parameter a)?

$P(X_1 = 1, X_2 = 2) = P(X_1 = 1) P(X_2 = 2 | X_1 = 1)$

$= \pi_1 q_{21} = \frac{1}{2} \cdot a = \frac{1}{2} a$

(b) (2 pts) What is the probability that you are well on the first day (state = 1) and then get a cold (state = 2) the next day, and that you feel high energy the first day and low energy the second day. That is, what is $P(X_1 = 1, X_2 = 2, Y_1 = H, Y_2 = L)$(Your final answer must involve numbers and the parameter a) ?

$P(X_1=1, X_2=2, Y_1=H, Y_2=L)$

$= P(X_1=1)\, P(Y_1=H \mid X_1=1)\, P(X_2=2 \mid X_1=1)\, P(Y_2=L \mid X_2=2)$

$= \pi_1 q_{2|1}\, e_1(H)\, e_2(L)$

$= \frac{1}{2} a(1-a) \cdot 1$

$= \frac{1}{2} a(1-a)$

22. (8 pts) **Clustering**
Recall that in $K-$means clustering we attempt to find K cluster centroids $\boldsymbol{\mu}_k \in \mathbb{R}^d, k \in \{1, \ldots, K\}$ such that the total distance between each datapoint and the nearest cluster centroid is minimized. In other words, we attempt to solve:

$$\min_{\{\boldsymbol{\mu}_k\},\{r_{nk}\}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2, \tag{1}$$

where N is the number of data points, r_{nk} is a binary variable that is 1 if sample n is assigned to cluster k and zero otherwise.

(a) (3 pts) Instead of holding the number of clusters K fixed, one can think of minimizing (1) over all of K and $\boldsymbol{\mu}_k$ and r_{nk}. Show that this is a bad idea. Specifically, what is the minimum possible value of (1) ? What values of K and $\boldsymbol{\mu}_k$ result in this value ?

If we don't put restrictions on K, $K = N$ would result in an objective value 0, and $\mu_n = x_n$ in this case.

(b) (2 pts) Recall that in one of the steps of the K-means algorithm, for a fixed assignment of each sample to one of the K clusters (r_{nk}), we compute the new cluster centroids $\boldsymbol{\mu}_k$ by minimizing $\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2$.
Compute the gradient of the above objective function with respect to $\boldsymbol{\mu}_k$. For reference, here is a useful identity:

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 \qquad \nabla f(\boldsymbol{x}) = 2\boldsymbol{x}$$

Define $J(\mu_k) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2$

$$\nabla J(\mu_k) = \sum_{n=1}^{N} -2 r_{nk}(x_n - \mu_k)$$

$$= -2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k)$$

(c) (1 pts) Set the gradient to zero and solve for $\boldsymbol{\mu}_k$. Show that the optimal $\boldsymbol{\mu}_k^*$ corresponds to the mean of the samples assigned to cluster k.

$$\nabla J(\mu_k) = -\sum_{n=1}^{N} 2 r_{nk}(x_n - \mu_k) = 0$$

$$\Rightarrow \quad \sum_{n=1}^{N} r_{nk}\mu_k = \sum_{n=1}^{N} r_{nk} x_n$$

$$\Rightarrow \quad \mu_k^* = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}}$$

↑

Mean of the samples assigned to cluster k

(d) (2 pts) We now consider clustering 1D data using K-means. We assume the number of clusters $K = 2$. You are given four instances: $(x_1, x_2, x_3, x_4) = (1, 10, 20, 9)$ where each $x_n \in \mathbb{R}, n \in \{1, 2, 3, 4\}$. The current estimates of r_{nk} is represented by the following matrix:

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Here entry (n, k) of this matrix is $r_{n,k}$.

Show the update for the cluster centroids μ_1, μ_2 (you do not need to simplify your answer).

$\mu_1 = x_1 = 1$

$\mu_2 = \frac{x_2 + x_3 + x_4}{3} = \frac{10 + 20 + 9}{3} = 13$

(Blank page provided for your work)

$(X^T y)^T \theta$

$(y^T - \theta^T X^T)(y - X\theta) + \lambda \theta^T \theta$

$y^T y - y^T X\theta - \theta^T X^T y + \theta^T X^T X \theta + \lambda \theta^T \theta$

$(y - X\theta) + \lambda\theta = 0$

$\theta =$

$-2X^T y + 2X^T X + 2\lambda\theta = 0$

$\hat{\theta} = \frac{X^T y - X^T X}{2\lambda}$

$= \frac{X^T (y - X)}{T}$