CM146 Final

PRASANNA SHREE KESAVA NARAYAN

TOTAL POINTS

98 / 100

QUESTION 1

1 Name and UID **2 / 2**

✓ **- 0 pts** Correct

QUESTION 2

True/False 20 pts

2.1 1-4 **8 / 8**

✓ **- 0 pts** Correct

2.2 5-9 **10 / 10**

✓ **- 0 pts** 9 arguably incorrect

2.3 10 **2 / 2**

✓ **- 0 pts** Correct

QUESTION 3

Multiple choice 32 pts

3.1 11 **4 / 4**

✓ **- 0 pts** Correct

3.2 12 **4 / 4**

✓ **- 0 pts** Correct

3.3 13 **2 / 4**

✓ **- 2 pts** Selected 2 correct answer

3.4 14 **4 / 4**

✓ **- 0 pts** Correct

3.5 15 **4 / 4**

✓ **- 0 pts** Correct

3.6 16 **4 / 4**

✓ **- 0 pts** Correct

3.7 17 **4 / 4**

✓ **- 0 pts** Correct

3.8 18 **4 / 4**

✓ **- 0 pts** Correct

QUESTION 4

4 19 **4 / 4**

✓ **- 0 pts** Correct

QUESTION 5

5 20 **6 / 6**

✓ **- 0 pts** Correct

QUESTION 6

21 6 pts

6.1 (a-d) performance **4 / 4**

✓ **- 0 pts** Correct

6.2 (e-f) Precision **2 / 2**

✓ **- 0 pts** Correct

QUESTION 7

Kernelized K-means 10 pts

7.1 Mean update as linear comb. of vectors **3 / 3**

✓ **- 0 pts** Correct

7.2 Squared distance as linear comb. of vectors **3 / 3**

✓ **- 0 pts** Correct

7.3 Point to centroid squared distance **4 / 4**

✓ **- 0 pts** Correct

QUESTION 8

Poisson Regression 10 pts

8.1 Log likelihood 4 / 4

✓ **- 0 pts** Correct

8.2 Gradient 6 / 6

✓ **- 0 pts** Correct

QUESTION 9

SVM 10 pts

9.1 Parallel to optimal 2 / 2

✓ **- 0 pts** Correct

9.2 Margin 2 / 2

✓ **- 0 pts** Correct

9.3 Optimal vector 2 / 2

✓ **- 0 pts** Correct

9.4 Optimal offset 2 / 2

✓ **- 0 pts** Correct

9.5 Prediction 2 / 2

✓ **- 0 pts** Correct

CM146: Introduction to Machine Learning **Winter 2019**

Final Exam

March 19^{th}, 2019

- Please do not open the exam unless you are instructed to do so.
- This is a closed book and closed notes exam.
- Everything you need in order to solve the problems is supplied in the body of this exam OR in a cheatsheet at the end of the exam.
- Mark your answers ON THE EXAM ITSELF. If you make a mess, clearly indicate your final answer (box it).
- For true/false questions, CIRCLE True OR False and provide a brief justification for full credit.
- Unless otherwise instructed, for multiple-choice questions, CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one) and provide a brief justification if the question asks for one.
- If you think something about a question is open to interpretation, feel free to ask the instructor or make a note on the exam.
- If you run out of room for your answer in the space provided, please use the blank pages at the end of the exam and indicate clearly that you've done so.
- DO NOT put answers on the back of any page of the exam.
- DO NOT detach ANY pages.
- You may use scratch paper if needed (provided at the end of the exam).
- You have 2 hours 45 minutes.
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**

Good Luck!

Legibly write your name and UID in the space provided below to earn 2 points.

Name: SHREE KESAVA NARAYAN PRASANNA

UID: 004 973 979

Name and UID		/2
True/False		/20
Multiple choice		/32
Short questions		/16
Poisson regression		/10
Kernelized K-means		/10
SVM		/10
Total		/100

1 True or False (20 pts)

Choose either True or False for each of the following statements. (If your answer is incorrect, partial credit may be given if your explanation is reasonable.)

1. (2 pts) After mapping the instances into a high dimensional space, a Perceptron may be able to achieve better classification performance on instances it was not able to classify before.

   [✓ True] False

   In the case that the sample data is not linearly separable, a non-linear transformation $\phi(\cdot)$ may be applied, mapping instances to higher dimensional space. In this new space, the data may indeed be linearly separable, and so the perception may perform better.

2. (2 pts) A neural network with all linear activation functions can learn non-linear decision boundaries.

   True [✓ False]

   A NN with all linear activation functions could be modelled by a single layer NN, with linear activation, ~~as~~ At every layer, a linear combination of outputs from previous layer in computed. without non-linear activation, the final output would just be a linear combination of the input layer, and so only a linear boundary would be learnt.

3. (2 pts) In the AdaBoost algorithm, at each iteration, we increase the weight for misclassified examples.

   [✓ True] False

   $w_{t+1} \propto (e^{-ya_t(x)}) w_t$. If the ~~#~~ example is misclassified, $-ya_t(x)$ is > 0, so $e^{-ya_t(x)} > 1$, so ~~w~~ the weight for that example is increased.

4. (2 pts) The training error of 1-Nearest Neighbor classifier is zero even if the dataset is not linearly separable.

   [✓ True] False

   To compute train error, each train example is classified and accuracy is computed. In 1-NN case, the neighbour closest to an ~~i~~ input train data sample is that data sample itself, so it will always be correctly classified, so train error ~~&~~ is 0.

5. (2 pts) Given a function $k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ for all $\boldsymbol{x}_i, \boldsymbol{x}_j$, k is a valid kernel function.

True [✓ False]

Counter example $K(\cdot,\cdot) = \|x_i - x_j\|_2^2$.

Mercer Theorem: $\begin{pmatrix} K(x_i, x_i) & K(x_j, x_i) \\ K(x_i, x_j) & K(x_j, x_j) \end{pmatrix}$ must be $\succeq 0$ for K to be a Kernel

But $\begin{pmatrix} 0 & \|x_i - x_j\|_2^2 \\ \|x_i - x_j\|_2^2 & 0 \end{pmatrix}$ has positive & negative eigenvalues so it is not $\succeq 0$.

So K is not a kernel

6. (2 pts) Ridge regression is <u>more</u> likely to overfit when we increase λ where λ is the non-negative weight for the regularization term.

True [✓ False]

$J(\theta) = l(\theta) + \lambda \|\theta\|_2^2$ → Ridge regression objective.

with increase in λ, ~~the~~ greater is the influence of $\|\theta\|_2^2$ on the objective, so an optimization algorithm would tend to decrease $\|\theta\|_2^2$ further, leading to more regularization and hence less overfitting.

7. (2 pts) The hard-margin SVM can <u>sometimes</u> fail to find a solution.

[✓ True] False

The hard-margin SVM ~~assumes~~ is ~~a~~ in-feasible when the data is not linearly separable.

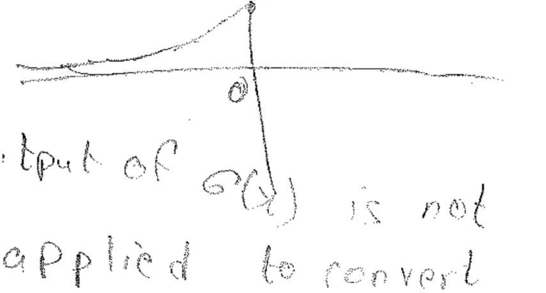
8. (2 pts) The minimum value of the K-means objective function can be zero for large enough K.

[✓ True] False

If K is N, the # of training examples, each example would be its own cluster, ~~and the~~ $J = \sum_n \sum_K r_{nk} \|x_n - \mu_k\|_2^2$ would be 0 as the distance of any point from its cluster center would be 0, as each point would be its own cluster center.

9. (2 pts) A sigmoid function, $\sigma(x) = 1/(1 + \exp(-x))$, can map the hidden layer output of a neural network to a Boolean/binary output.

True [✓ False]



~~sign~~ $\sigma(x)$ $0 < \sigma(x) \leq 1$. $\sigma(x)$ alone maps ~~is~~ hidden layer output to a real value between 0 & 1. So, the output of $\sigma(x)$ is not binary. Some thresholding may be applied to convert the output to 0/1 depending on the threshold, but that is separate from $\sigma(x)$ itself.

10. (2 pts) Given a sentence consisting of T words $(y_1, y_2, \ldots, y_T)$, we would like to use a hidden Markov model (HMM) to predict parts of speech (POS) tags x_t for each word $y_t, t \in \{1, \ldots, T\}$. We have $x_t \in \{1, \ldots, A\}, y_t \in \{1, \ldots, B\}$ for $t \in \{1, \ldots, T\}$ where the total number of possible words is B and the total number of POS tags is A. You use the Viterbi algorithm to compute the most probable sequence $(x_1, \ldots, x_T)$. The computational complexity of the algorithm scales as $\mathcal{O}(B^2T)$.

True [✓ False]

The computational complexity would scale as $O(A^2T)$, as A: # of hidden states, and the computation complexity depends on # of hidden states, not on # of emission symbols.

Multiple choice (32 pts)

CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one)

11. (4 pts) Which of these is an example of an unsupervised learning problem ?

    (a) From emails labeled as spam/not-spam, learn to predict if an email is spam.

    (b) From documents labeled by their topic, learn to classify a document into topics.

    (c) From images of handwritten digits labeled with the digit, learn a handwritten digit classifier.

    [✓ (d)] From a set of documents, group documents according to their topics.

12. (4 pts) Your classifier has substantially higher test error than training error. Which of the following statements could explain this observation?

    [✓ (a)] The hypothesis space from which the classifier was chosen is too complex.

    (b) The hypothesis space from which the classifier was chosen is too simple.

    [✓ (c)] The distribution of test data differs from the distribution of training data.

    [✓ (d)] The size of the training dataset is too small.

13. (4 pts) You are given a training dataset for a binary classification problem: $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ where $(x_n, y_n), n \in \{1, \ldots, N\}$ is an instance-label pair and $y_n \in \{0, 1\}$ where 1 refers to the positive class. Your classifier predicts 1 with probability $0 < \theta < \frac{1}{2}$ and 0 otherwise, independent of the features of an instance. Assume the sample size N is large. Which of following statements is true ?

    [✓ (a)] The sensitivity of the classifier is θ.

    (b) The specificity of the classifier is θ.

    [✓ (c)] The false positive rate of the classifier is θ.

    (d) The ROC curve for this classifier is the diagonal line.

14. (4 pts) Given the same training data consisting of N instances and D features, we fit linear regression and obtain the optimal parameters $\boldsymbol{\theta}_{OLS}$. We also fit ridge regression with regularization parameter $\lambda > 0$ and obtain the optimal parameters $\boldsymbol{\theta}_{Ridge}$. Let $RSS(\boldsymbol{\theta})$ denote the residual sum of squares cost function evaluated on the training set for the model associated with the parameter $\boldsymbol{\theta}$. Which of the following will always hold ?

[✓ (a)] $RSS(\boldsymbol{\theta}_{Ridge}) \geq RSS(\boldsymbol{\theta}_{OLS})$

(b) $RSS(\boldsymbol{\theta}_{Ridge}) \leq RSS(\boldsymbol{\theta}_{OLS})$

(c) $RSS(\boldsymbol{\theta}_{Ridge}) = RSS(\boldsymbol{\theta}_{OLS})$

[✓ (d)] $RSS(\boldsymbol{\theta}_{OLS}) \geq 0$.

15. (4 pts) Let $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ be the design matrix with each row corresponding to the features of an example and $\boldsymbol{y} \in \mathbb{R}^N$ be a vector of all the labels. The OLS solution is $\boldsymbol{\theta}_{OLD} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$. Given a new test data point $\boldsymbol{x}$, our prediction for this data point is given by $\hat{y}_{OLD} = \boldsymbol{\theta}_{OLD}{}^{\mathrm{T}}\boldsymbol{x}$. We then scale each feature in the training and the test data by 2 and compute the OLS solution $\boldsymbol{\theta}_{NEW}$ to make our prediction $\hat{y}_{NEW} = \boldsymbol{\theta}_{NEW}{}^{\mathrm{T}}\boldsymbol{x}$. What is the relation between $\hat{y}_{NEW}$ and $\hat{y}_{OLD}$?

(a) $\hat{y}_{NEW} = 2\hat{y}_{OLD}$

(b) $\hat{y}_{NEW} = 4\hat{y}_{OLD}$

(c) $\hat{y}_{NEW} = \frac{1}{2}\hat{y}_{OLD}$

[✓ (d)] $\hat{y}_{NEW} = \hat{y}_{OLD}$

16. (4 pts) Random variables (X_1, X_2, X_3, X_4) are distributed according to a Markov model. Which of the following statements is true of the distributions of these random variables?

(a) $P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2)P(x_3)P(x_4)$

[✓ (b)] $P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)P(x_4|x_1, x_2, x_3)$

[✓ (c)] $P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)$

[✓ (d)] $P(x_4|x_1, x_2, x_3) = P(x_4|x_3)$

17. (4 pts) We would like to run PCA on a dataset with 100 samples and five features. The eigenvalues of the covariance matrix are $(80, 15, 5, 0, 0)$. What is the minimum number K of principal components needed so that the transformed K features explain at least 90% of the variance?

   (a) 1
   [✓ (b) 2]
   (c) 3
   (d) 4

18. (4 pts) We would like to apply the EM algorithm to estimate the parameters of Gaussian Mixture Models (GMMs). Which of the following are true of the EM algorithm applied to GMMs?

   [✓ (a) In the E-step, we compute the probability that a data point is drawn from each mixture component.]
   (b) In the E-step, each data point is assigned to one of the mixture components.
   [✓ (c) In the M-step, we update the mixture weights of the GMM.]
   [✓ (d) In the M-step, we update the mean and covariance matrix of each mixture component.]

18. Assumed for (a) that the probability being computed is $p(z_n = k \mid x_n)$

(b) is not chosen as each data point is only softly assigned to each cluster (mixture component).

Short Answer Questions (16 pts)

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

19. (4 pts) Describe the difference between *maximum likelihood estimation* (MLE) and *maximum a posteriori estimation* (MAP), and state under what condition MAP is equivalent to MLE?

Given a dataset D, we want to find θ

MLE finds that θ that maximizes $P(D|\theta)$

MAP finds that θ that maximizes $P(\theta|D)$

Note that $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$. MAP & MLE would yield the same results if the prior distribution of θ is uniform ($P(\theta)$ = constant over all θ)

20. (6 pts) Given vectors x and z in R^3, define the kernel $K_\beta(x;z) = (\beta + x \cdot z)^2$ for any value $\beta > 0$. Find the corresponding feature map $\phi_\beta(\cdot)$.

$$(\beta + x \cdot z)^2 = (\beta + x_1z_1 + x_2z_2 + x_3z_3)^2 = \beta^2 + 2\beta x_1z_1 + 2\beta x_2z_2 + 2\beta x_3z_3 + x_1^2z_1^2 + 2x_1x_2z_1z_2 + 2x_1x_3z_1z_3 + x_2^2z_2^2 + 2x_2x_3z_2z_3 + x_3^2z_2^2$$

$$= \begin{pmatrix} \beta \\ \sqrt{2\beta}\,x_1 \\ \sqrt{2\beta}\,x_2 \\ \sqrt{2\beta}\,x_3 \\ x_1^2 \\ x_2^2 \\ x_3^2 \\ \sqrt{2}\,x_1x_2 \\ \sqrt{2}\,x_2x_3 \\ \sqrt{2}\,x_1x_3 \end{pmatrix}^T \begin{pmatrix} \beta \\ \sqrt{2\beta}\,z_1 \\ \sqrt{2\beta}\,z_2 \\ \sqrt{2\beta}\,z_3 \\ z_1^2 \\ z_2^2 \\ z_3^2 \\ \sqrt{2}\,z_1z_2 \\ \sqrt{2}\,z_2z_3 \\ \sqrt{2}\,z_1z_3 \end{pmatrix} \qquad \phi_\beta(x) = \begin{pmatrix} \beta \\ \sqrt{2\beta}\,x_1 \\ \sqrt{2\beta}\,x_2 \\ \sqrt{2\beta}\,x_3 \\ x_1^2 \\ x_2^2 \\ x_3^2 \\ \sqrt{2}\,x_1x_2 \\ \sqrt{2}\,x_2x_3 \\ \sqrt{2}\,x_1x_3 \end{pmatrix}$$

21. (6 pts) We are given the confusion matrix for a binary classifier.

	Actual class	Negative	Positive
Predicted class			
Negative		80	50
Positive		20	50

Compute the following quantities related to its accuracy:

(a) (1 pts) True positives

50

(b) (1 pts) False positives

20

(c) (1 pts) Recall

Recall = $\frac{TP}{P} = \frac{50}{100} = 0.5$

(d) (1 pts) False positive rate

$= \frac{FP}{N} = \frac{20}{100} = 0.2$

(e) (1 pts) Precision

$$= \frac{TP}{FP+TP} = \frac{50}{20+50} = \frac{5}{7}$$

(f) (1 pts) Accuracy

$$= \frac{TP+TN}{P+N} = \frac{50+80}{200} = 0.65 = \frac{13}{20}$$

22 Kernelized K-means (10 pts)

K-means with Euclidean distance metric assumes that each pair of clusters is linearly separable. This may not be the case. We have seen that we can use kernels to obtain a non-linear version of an algorithm that is linear by nature and K-means is no exception. Recall that there are two main aspects of kernelized algorithms: (i) the solution is expressed as a linear combination of training examples, (ii) the algorithm relies only on inner products between data points rather than their explicit representation. We will show that these two aspects can be satisfied in K-means.

1. (3 pts) Let z_{nk} be an indicator that is equal to 1 if the x_n is currently assigned to the k^{th} cluster and 0 otherwise ($1 \leq n \leq N$ and $1 \leq k \leq K$). Show that the k^{th} cluster center $\boldsymbol{\mu}_k$ can be updated as $\sum_{n=1}^{N} \alpha_{nk}\boldsymbol{x}_n$. Specifically, show how α_{nk} can be computed given all z's.

We know that

$$\mu_K = \frac{\sum_n z_{nk} x_n}{\sum_n z_{nK}} \qquad \text{Let } \boxed{\alpha_{nK} = \frac{z_{nK}}{\sum_n z_{nK}}}$$

$\therefore$, α_{nK} can be computed given all z's.

Then $$\mu_K = \frac{z_{1K}}{\sum_n z_{nK}} x_1 + \frac{z_{2K} x_2}{\sum_n z_{nK}} + \cdots + \frac{z_{NK} x_N}{\sum_n z_{nK}} = \alpha_{1K} x_1 + \cdots + \alpha_{NK} x_N = \sum_{n=1}^{N} \alpha_{nK} x_n$$

2. (3 pts) Given two data points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, show that the square distance $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2$ can be computed using only (linear combinations of) inner products.

$$\|x_1 - x_2\|_2^2 = (x_1 - x_2)^T (x_1 - x_2) = (x_1^T - x_2^T)(x_1 - x_2)$$
$$= \underbrace{x_1^T x_1 - 2x_1^T x_2 + x_2^T x_2}_{}$$

This is a linear combination of inner products.

$(x_1^T x_2 = x_2^T x_1)$

3. (4 pts) Given the results of the above two parts, show how to compute the squared distance $\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2$ using only (linear combinations of) inner products between the data points. $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (You can leave your answer in terms of α_{nk} and inner product of $\boldsymbol{x}_n$ and $\boldsymbol{x}_k$.

$$\|x_n - \mu_K\|^2 = x_n^T x_n - 2 x_n^T \mu_K + \mu_K^T \mu_K$$

$$\mu_K = \sum_{m=1}^{N} \alpha_{mK} x_m$$

$$\therefore = x_n^T x_n - 2 x_n^T \left(\sum_{m=1}^{N} \alpha_{mK} x_m\right) + \left(\sum_{m=1}^{N} \alpha_{mK} x_m\right)^T \left(\sum_{m=1}^{N} \alpha_{mK} x_m\right)$$

$$= x_n^T x_n - 2 \sum_{m=1}^{N} \left(\alpha_{mK} x_n^T x_m\right) + \sum_{m=1}^{N} \sum_{j=1}^{N} \alpha_{mK} \alpha_{jK} x_m^T x_j$$

Note: This means that given a kernel K, we can run Lloyd's algorithm. We begin with some initial data points as centers and use the answer to part c) to find the closest center for each data point, giving us the initial z_{nk}'s. We then repeatedly use the answer to part a) to reassign the cluster centers and use the answer to part c) to reassign points to centers and update the z_{nk}'s.

23 Poisson Regression (10 pts)

We want to predict the number of user clicks on a website. The number of clicks takes on values in $\{0, 1, 2, \ldots\}$. In class, we showed how linear regression (ordinary least squares) can be interpreted as a probabilistic model where the output is real-valued. Logistic regression is a probabilistic model where the output takes values in $\{0, 1\}$. In this problem, we want to model outputs in $\{0, 1, 2, \ldots\}$.

In this example, we have a training set $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ where $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \{0, 1, 2, \ldots\}$. Now we model our target y_n as distributed according to a Poisson distribution. Specifically

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

1. (4 pts) Write the log likelihood of the parameters $l(\boldsymbol{\theta})$. Express your answer in terms of $y_n, \boldsymbol{x}_n, \boldsymbol{\theta}$.

$$L(\theta) = \prod_{n=1}^{N} P(y_n|x_n;\theta) \qquad \ell(\theta) = \log L(\theta) = \log \prod_{n=1}^{N} P(y_n|x_n;\theta),$$

$$= \sum_{n=1}^{N} \log P(y_n|x_n;\theta)$$

$$= \sum_n \log\left(\frac{1}{y_n!} e^{(y_n \theta^T x_n)} e^{-e^{\theta^T x_n}}\right) = \sum_n \left(y_n \theta^T x_n - e^{\theta^T x_n} - \log y_n!\right)$$

$$\therefore \ell(\theta) = \sum_{n=1}^{N} \left(y_n \theta^T x_n - e^{\theta^T x_n} - \log(y_n!)\right)$$

2. (6 pts) Show that the gradient can be written in the form $\nabla l(\boldsymbol{\theta}) = \sum_{n=1}^{N} \epsilon_n \boldsymbol{x}_n$ for some ϵ_n. Write ϵ_n in terms of $\boldsymbol{x}_n, y_n$ and $\boldsymbol{\theta}$.

$$\nabla l(\theta) = \nabla \sum_{n=1}^{N} \left(y_n \theta^T x_n - e^{\theta^T x_n} - \log(y_n!) \right)$$

$$\nabla_\theta y_n \theta^T x_n = y_n x_n$$

$$\nabla_\theta e^{\theta^T x_n} = \frac{\partial e^{\theta^T x_n}}{\partial(\theta^T x_n)} \nabla(\theta^T x_n) = e^{\theta^T x_n} x_n$$

$$\nabla_\theta \log(y_n!) = 0$$

$$\nabla l(\theta) = \sum_{n=1}^{N} y_n x_n - e^{\theta^T x_n} x_n$$

$$= \sum_{n=1}^{N} (y_n - e^{\theta^T x_n}) x_n$$

$$= \underline{\sum_{n=1}^{N} \epsilon_n x_n}, \quad \text{where} \quad \underline{\epsilon_n = y_n - e^{\theta^T x_n}}$$

24 SVM (10 pts)

We are attempting to use hard-margin SVM to solve a binary classification problem given a dataset $\mathcal{D}$ that has two samples $\{(x_1, y_1), (x_2, y_2)\}$ with $x_i \in R$ and $y_i \in \{-1, +1\}$, $(x_1 = 0, y_1 = -1)$ and $(x_2 = \sqrt{2}, y_2 = 1)$. To obtain a non-linear classifier, consider mapping the data points by $\phi(x) = [1, \sqrt{2}x, x^2]^T$ to a 3-dimensional space. The hard-margin SVM has the form

$$\begin{aligned} min_{\boldsymbol{w},b} \quad & \frac{1}{2}||\boldsymbol{w}||_2^2 \\ \text{s.t.} \quad & y_1(\boldsymbol{w}^T\phi(x_1) + b) \geq 1 \\ & y_2(\boldsymbol{w}^T\phi(x_2) + b) \geq 1 \end{aligned} \tag{1}$$

1. (2 pts) Write a vector that is parallel to the optimal vector $\boldsymbol{w}^*$ and justify your answer.

Lagrangian $L = \frac{1}{2}\|w\|_2^2 + \alpha_1(1 - y_1 w^T\phi(x_1) - y_1 b) + \alpha_2(1 - y_2 w^T\phi(x_2) - y_2 b)$

Dual: $\max_{\alpha_1, \alpha_2} g(\alpha_1, \alpha_2)$, $g(\alpha_1, \alpha_2) = \min_{w,b} L$ so, $\nabla_w L = w - \alpha_1 y_1 \phi(x_1) - \alpha_2 y_2 \phi(x_2) = 0$

$\Rightarrow w = \alpha_1(-1)\phi(0) + \alpha_2(1)\phi(\sqrt{2}) = -\alpha_1\begin{bmatrix}1\\0\\0\end{bmatrix} + \alpha_2\begin{bmatrix}1\\2\\2\end{bmatrix}$. Also $\frac{\partial L}{\partial b} = 0$

$\Rightarrow -y_1\alpha_1 - y_2\alpha_2 = 0 \Rightarrow \alpha_1 - \alpha_2 = 0 \Rightarrow \alpha_1 = \alpha_2 \therefore w^* = \cancel{\alpha}\, \alpha_1\begin{bmatrix}0\\2\\2\end{bmatrix}$

$\therefore \begin{bmatrix}0\\2\\2\end{bmatrix}$ is parallel to w^* as $w^* = \alpha_1\begin{bmatrix}0\\2\\2\end{bmatrix}$, α_1 is a scalar, $\triangleright\ \alpha_1 \geq 0$

2. (2 pts) Write down the value of the margin achieved by the optimal $\boldsymbol{w}^*$.

(continuing 1) Another explanation is that: $\phi(x_1) = \begin{bmatrix}1\\0\\0\end{bmatrix}$, $\phi(x_2) = \begin{bmatrix}1\\2\\2\end{bmatrix}$ $\phi(x_2) - \phi(x_1) = \begin{bmatrix}0\\2\\2\end{bmatrix}$. The hyperplane would have a normal parallel to the vector joining $\phi(x_1)$ & $\phi(x_2)$.

To maximize margin, want to maximize the minimum distance of any point to the ~~&~~ hyperplane. With two points, this is achieved when the hyperplane ~~&~~ is equidistant from the two points. ↳ If ~~*~~ it weren't, it would be closer to one point than the other, and the min distance would be less than the halfway length. So max margin = halfway length

$= \|\phi(x_1) - \phi(x_2)\|_2 / 2 = \left\|\begin{bmatrix}0\\2\\2\end{bmatrix}\right\|_2 / 2 = \boxed{\sqrt{2}}$

3. (2 pts) Solve for $\boldsymbol{w}^*$ using the fact that the margin is equal to $\frac{1}{\|\boldsymbol{w}^*\|_2}$.

$$\underbrace{\sqrt{2}}_{\text{Margin}} = \frac{1}{\|w^*\|_2} \Rightarrow \|w^*\|_2 = \frac{1}{\sqrt{2}} \Rightarrow \alpha \left\| \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix} \right\|_2 = \frac{1}{\sqrt{2}}$$

$$\Rightarrow \alpha (2\sqrt{2}) = \frac{1}{\sqrt{2}} \Rightarrow \underline{\underline{\alpha = \frac{1}{4}}}$$

$$\therefore w^* = \frac{1}{4} \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix} = \underline{\underline{\begin{bmatrix} 0 \\ 1/2 \\ 1/2 \end{bmatrix}}}$$

4. (2 pts) Solve for b^* and write down the decision boundary $f(x) = \boldsymbol{w}^{*T}\boldsymbol{\phi}(x) + b^*$.

We must satisfy the constraints

$$y_1(w^{*T}\phi(x_1) + b^*) \geq 1$$

$$y_2(w^{*T}\phi(x_2) + b^*) \geq 1$$

$$\Rightarrow -1\left(\begin{bmatrix} 0 \\ 1/2 \\ 1/2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + b^*\right) \geq 1$$

$$\Rightarrow -b^* \geq 1 \Rightarrow b^* \leq -1$$

&

$$1\left(\begin{bmatrix} 0 \\ 1/2 \\ 1/2 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} + b^*\right) \geq 1 \Rightarrow 2 + b^* \geq 1$$

$$\Rightarrow b^* \geq -1$$

Now $b^* \leq -1$ & $b^* \geq -1$

$$\Rightarrow \underline{\underline{b^* = -1}}$$

$$\therefore f(x) = \underline{\underline{\begin{bmatrix} 0 \\ 1/2 \\ 1/2 \end{bmatrix}^T \begin{bmatrix} 1 \\ \sqrt{2}\,x \\ x^2 \end{bmatrix} - 1}} = \underline{\underline{\frac{x}{\sqrt{2}} + \frac{x^2}{2} - 1}}$$

5. (2 pts) Given a new data point x_{new}, what is the prediction of the label for x_{new} using the above parameters?

$$\text{sign } f(x_{new}) = \text{sign}\left(\frac{x_{new}}{\sqrt{2}} + \frac{x_{new}^2}{2} - 1\right)$$

Identities

Probability density/mass functions for some distributions

$$\begin{aligned}
\text{Normal} &: \; P(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
\text{Multinomial} &: \; P(\boldsymbol{x};\boldsymbol{\pi}) = \prod_{k=1}^{K} {\pi_k}^{x_k} \\
&\quad \boldsymbol{x} \text{ is a length } K \text{ vector with exactly one entry equal to 1} \\
&\quad \text{and all other entries equal to 0} \\
\text{Poisson} &: \; P(x;\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}
\end{aligned}$$

Matrix calculus

Here $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{b} \in \mathbb{R}^n, \boldsymbol{A} \in \mathbb{R}^{n\times n}$. $\boldsymbol{A}$ is symmetric.

$$\begin{aligned}
\nabla \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} &= 2\boldsymbol{A}\boldsymbol{x} \\
\nabla \boldsymbol{b}^{\mathrm{T}} \boldsymbol{x} &= \boldsymbol{b}
\end{aligned}$$

You may use this page for scratch space.

You may use this page for scratch space.