Final

Graded

Student

BRANDON LO

Total Points

79 / 100 pts

Question 1

True or False **22** / 24 pts

1.1 **Gradient Descent vs. Closed Form** **2** / 2 pts

✔ **– 0 pts Correct**: True

1.2 **AdaBoost Contributions** **2** / 2 pts

✔ **– 0 pts Correct**: True

1.3 **Optimal K-means** **2** / 2 pts

✔ **– 0 pts Correct**: False

1.4 **Variance in PCA** **2** / 2 pts

✔ **– 0 pts Correct**: True

1.5 **SVM Margin** **2** / 2 pts

✔ **– 0 pts Correct**: False

1.6 **Multi-class Classification** **2** / 2 pts

✔ **– 0 pts Correct**: False

1.7 **Residual Sum of Squares** **0** / 2 pts

✔ **– 2 pts Incorrect**: True

1.8 **Ridge Regression** **2** / 2 pts

✔ **– 0 pts Correct**: True

1.9 **Logistic Regression** **2** / 2 pts

✔ **– 0 pts Correct**: False

1.10 **Convexity of RSS** **2** / 2 pts

✔ **– 0 pts Correct**: True

1.11 **SGD vs. Batch GD** **2** / 2 pts

✔ **– 0 pts Correct**: False

1.12 **Hidden Markov Model** **2** / 2 pts

✔ **– 0 pts Correct**: False

Question 2

Multiple choice **23** / 30 pts

2.1 **Zero Training Error** **3** / 3 pts

✔ **– 0 pts Correct**: Selected options A, B, C

2.2 **Ridge Regression vs. OLS** **3** / 3 pts

✔ **– 0 pts Correct**: Selected options A, D

2.3 **Overfitting** **3** / 3 pts

✔ **– 0 pts Correct**: Selected options A, B

2.4 **Covariance Matrix** **3** / 3 pts

✔ **– 0 pts Correct**: Selected option D

2.5 **Slack Variables** **3** / 3 pts

✔ **– 0 pts Correct**: Selected option A

2.6 **Kernel Function** **0** / 3 pts

✔ **– 3 pts Incorrect**: One or more options incorrectly selected/unselected

2.7 **Markov Model** **3** / 3 pts

✔ **– 0 pts Correct**: Selected options B, C, D

2.8 **Gaussian Mixture Model** **1** / 3 pts

✔ **– 2 pts Partial**: Two options incorrectly selected/unselected

2.9 **Derivative of Swish** **3** / 3 pts

✔ **– 0 pts Correct**: Selected option D

2.10 **Activation Function Approximation** **1** / 3 pts

✔ **– 2 pts Partial**: Two options incorrectly selected/unselected

Question 3

Weighted linear regression — **5** / 14 pts

3.1 **Linearizing Cost Function** — **3** / 3 pts

✔ **– 0 pts** Correct

3.2 **Optimal Value** — **1** / 4 pts

✔ **– 3 pts** Major error

3.3 **Normal Distribution** — **1** / 3 pts

✔ **– 2 pts** Wrote the log probability of a sample instead of log likelihood

3.4 **Maximum Likelihood** — **0** / 4 pts

✔ **– 4 pts** Incorrect

Question 4

SVM — **10** / 10 pts

4.1 **Parallel Vector** — **2** / 2 pts

✔ **– 0 pts** Correct

4.2 **Margin** — **2** / 2 pts

✔ **– 0 pts** Correct

4.3 **Optimal Hyperplane** — **2** / 2 pts

✔ **– 0 pts** Correct

4.4 **Decision Boundary** — **2** / 2 pts

✔ **– 0 pts** Correct

4.5 **New Prediction** — **2** / 2 pts

✔ **– 0 pts** Correct

Question 5

Kernelized K-means — **7** / 10 pts

5.1 **Centroid Update** — **3** / 3 pts

✔ **– 0 pts** Correct

5.2 **Square Distance** — **2** / 3 pts

✔ **– 1 pt** Correct approach but incorrect final answer

5.3 **Kernelized Update** — **2** / 4 pts

✔ **– 2 pts** Correct approach but incorrect final answer

Question 6

Naive Bayes — **10** / 10 pts

6.1 **Modeling Assumption** — **2** / 2 pts

✔ **– 0 pts** **Correct**: Selected options B, C

6.2 **Number of Parameters** — **2** / 2 pts

✔ **– 0 pts** **Correct**: Selected option A

6.3 **Maximum Likelihood Estimate** — **2** / 2 pts

✔ **– 0 pts** **Correct**: Selected option B

6.4 **Probability of Word** — **2** / 2 pts

✔ **– 0 pts** **Correct**: Selected option D

6.5 **Classifying New Document** — **2** / 2 pts

✔ **– 0 pts** **Correct**: Selected option B

Question 7

Name and UID — **2** / 2 pts

✔ **– 0 pts** Correct

CM146: Introduction to Machine Learning **Winter 2024**

Final Exam

March 19, 2024

- This is an open book, open notes exam.
- Electronic devices (laptops, tablets, phones, calculators) SHOULD BE TURNED OFF for the duration of the exam.
- This exam has a total of **19 pages** including the cover sheet.
- Please WRITE YOUR NAME AND UID ON EACH PAGE OF THE EXAM.
- Mark your answers ON THE EXAM ITSELF in the space provided.
- DO NOT put answers on the back of any page of the exam.
- DO NOT detach ANY pages.
- For multiple-choice questions, CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one).
- Useful formulas and scratch space are provided at the end of the exam.
- You have **2 hours 45 minutes**.

Good Luck!

Legibly write your name and UID in the space provided below.

Name: Brandon Lo

UID: 105753560

Name and UID		/2
True/False		/24
Multiple choice		/30
Weighted linear regression		/14
SVM		/10
Kernelized K-means		/10
Naive Bayes		/10
Total		/100

Brandon Lo
105753560

1 True or False (24 pts)

Choose either True or False for each of the following statements.

1. (2 pts) We are attempting to fit a linear regression model with a large number of features. In this setting, gradient descent is preferred to the closed-form solution for computational efficiency.

   True False

2. (2 pts) In AdaBoost, a weak learner learned at round t misclassifies 20 training points while a weak learner at round $t+1$ misclassifies 10 points. The weak learner from round $t+1$ can have a lower contribution than the one from round t ($\beta_{t+1} < \beta_t$).

   True False

3. (2 pts) The optimal value of the K-means cost function increases with increasing K.

   True False

   It is a summation, so based of slides may be true but if mean, loss number could decrease

4. (2 pts) You run PCA on a dataset with five input features. The eigenvalues of the covariance matrix are $(20, 10, 10, 5, 5)$. The first principal component (PC) explains 40% of the variance.

   True False

   $\frac{20}{50} = 0.4$

5. (2 pts) Removing a data point from the training dataset will always increase the margin of a support vector machine (SVM).

   True False

   If a data point that's not the closest to the separation line is removed, should be no change to margin.

6. (2 pts) When performing multi-class classification for 10 classes, a one-versus-rest approach requires training more binary classifiers than a one-versus-one approach.

   True False

Brandon L
10575356

7. (2 pts) We are using linear regression to predict height from weight and age. The optimal value of the residual sum of squares (RSS) will change if we measure weight in pounds instead of kilograms for each instance.

True False

8. (2 pts) We fit a ridge regression model on a dataset with 100 instances and 1,000 features. The solution to ridge regression is always unique for any $\lambda > 0$.

True False

9. (2 pts) To predict y from $\boldsymbol{x}$ where $y \in \{0, 1\}, \boldsymbol{x} \in \mathbb{R}^D$, we transform $\boldsymbol{x}$ by a function $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{C}\boldsymbol{x} + \boldsymbol{d}$ where $\boldsymbol{C}$ is a $M \times D$ matrix and $\boldsymbol{d} \in \mathbb{R}^M$. A logistic regression model with $\boldsymbol{\phi}(\boldsymbol{x})$ as input can learn a non-linear decision boundary.

True False

10. (2 pts) To learn a non-linear model to predict y from $\boldsymbol{x}$ (where $y \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^D$), we first transform $\boldsymbol{x}$ by a function $\boldsymbol{\phi}(\boldsymbol{x})$. The residual sum of squares (RSS) cost function for a linear regression model with $\boldsymbol{\phi}(\boldsymbol{x})$ as input is convex.

True False

11. (2 pts) Stochastic gradient descent is guaranteed to converge to the minimum of a convex function faster than batch gradient descent (faster here refers to runtime).

True False

Batch gradient descent computes exact while stochastic computes estimate, Batch can be faster with different parameters than stochastic

12. (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

Brandon L
1057535

2 Multiple choice (30 pts)

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

1. (3 pts) Which of the following methods can achieve zero training error on any linearly separable dataset?

   (a) Hard-margin SVM ← linearly separate
   (b) Logistic regression ←
   (c) Neural network ← can create linear / non-linear bounds
   (d) 3-Nearest Neighbor ✗

✓ 2. (3 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) $RSS(\boldsymbol{\theta}_{OLS}) \geq 0$.

3. (3 pts) Which setting(s) increase the chance of over-fitting ?

   (a) Increasing the number of features
   (b) Increasing the complexity of the hypothesis space
   (c) Increasing the value of the regularization hyperparameter
   (d) Increasing the number of training examples ✗

would want to reduce complexity and extraneous features to reduce over-fitting

✓ 4. (3 pts) Let $\lambda_1 > \lambda_2 > \ldots > \lambda_D$ be the eigenvalues of a sample covariance matrix $\boldsymbol{C}$ over D features. The solution to the optimization problem $max_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x}$ is.

   (a) λ_1
   (b) λ_D ← solution to min
   (c) 0
   (d) ∞

✗ can be any vector which is arbitrarily large

Brandon Le
1057535

5. (3 pts) Let ξ_n denote the slack variable associated with training instance n for a soft-margin SVM. Which of the following is always true?

(a) $\xi_n \geq 0$
(b) $\xi_n \geq 1$
(c) If data point n is a support vector, $\xi_n > 0$.
(d) If data point n is a support vector, $\xi_n > 1$.

$\xi_n \geq 0$ due to the way we incorporate it into the optimization problem

6. (3 pts) Given a valid kernel function $k_1(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}_1(\boldsymbol{x})^T \boldsymbol{\phi}_1(\boldsymbol{x}')$ where $\boldsymbol{\phi}_1 : \mathbb{R}^D \to \mathbb{R}^{M_1}$, what is the feature transformation $\boldsymbol{\phi}(\boldsymbol{x})$ corresponding to a new kernel $k(\boldsymbol{x}, \boldsymbol{x}') = 3k_1(\boldsymbol{x}, \boldsymbol{x}') + 4$.

(a) $\boldsymbol{\phi}(\boldsymbol{x}) = \sqrt{3}\boldsymbol{\phi}_1(\boldsymbol{x})$
(b) $\boldsymbol{\phi}(\boldsymbol{x}) = \sqrt{3\boldsymbol{\phi}_1(\boldsymbol{x})}$
(c) $\boldsymbol{\phi}(\boldsymbol{x}) = [\sqrt{3}\boldsymbol{\phi}_1(\boldsymbol{x}), 2]$
(d) $\boldsymbol{\phi}(\boldsymbol{x}) = [\sqrt{3\boldsymbol{\phi}_1(\boldsymbol{x})}, 2]$

$\sqrt{3\phi_1(x) + 4}$

✓ 7. (3 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) $P(x_4|x_1, x_2, x_3) = P(x_4|x_3)$

8. (3 pts) The Gaussian Mixture Model (GMM) models each data point $\{\boldsymbol{x}_n \in \mathbb{R}^D\}, n \in \{1, \ldots, N\}$ as arising i.i.d. from a mixture of K Gaussian distributions: $\boldsymbol{x}_n \sim \sum_{k=1}^{K} \omega_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. The EM algorithm applied to Gaussian Mixture Models (GMMs) reduces to the K-means algorithm under which of the following conditions ? (Here $\boldsymbol{I}_D$ denotes an identity matrix of D dimensions).

(a) $\omega_k = \frac{1}{K}$, $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}_D$
(b) $\omega_k = \frac{1}{K}$, $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}_D$, $\sigma^2 \to 0$
(c) $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}_D$, $\sigma^2 \to 0$
(d) $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}_D$

ω_k should be $\frac{1}{k}$ because it's a mean calculation.

Multiplying covariance by identity matrix makes sense

Brandon L
1057535

The deep learning community has explored the effectiveness of different activation functions for learning in neural networks. Consider the following activation function : $f(x) = x\sigma(\beta x)$ where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function and β is a hyperparameter.

9. (3 pts) What is the derivative $f'(x)$ (for a fixed value of β)?

   (a) $\sigma(\beta x)$
   (b) $\beta x\sigma(\beta x)(1-\sigma(\beta x))$
   (c) $\sigma(\beta x)+\beta\sigma(\beta x)(1-\sigma(\beta x))$
   (d) $\sigma(\beta x)+\beta x\sigma(\beta x)(1-\sigma(\beta x))$

$1\cdot\sigma(\beta x) + x\,\sigma'(\beta x)$

$\sigma(\beta x) + x\left(\sigma(\beta x)(1-\sigma(\beta x))\right)$

10. (3 pts) Which of the following common activation functions does $f(x)$ approximate as $\beta \to \infty$?

    (a) Sigmoid
    (b) tanh
    (c) ReLU
    (d) Linear

$x \to \infty$

$\frac{1}{1+\frac{1}{e^x}}$ $\qquad \frac{1}{e^x} \to 0$

$\frac{1}{1} = 1$

Brandon Lo
1057535

3 Weighted linear regression (14 points)

You are peforming a study to understand how ocean temperature y_n in a location n is affected by several features $\boldsymbol{x}_n$. You aim to fit a linear regression model of ocean temperature as a function of the features. However, the sensors that measure temperature are of varying accuracy. To account for this, you have a weight w_n associated with each location where larger values of w_n indicate that you trust this measurement more.

To learn a weighted linear regression model, you want to find parameters $\boldsymbol{\theta}$ that minimize the cost function: $J(\boldsymbol{\theta}) = \sum_{n=1}^{N} w_n(y_n - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_n)^2$ where $w_n > 0$, $\boldsymbol{x}_n \in \mathbb{R}^{D+1}, \boldsymbol{\theta} \in \mathbb{R}^{D+1}$. $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathrm{T}} \end{pmatrix} \in \mathbb{R}^{N\times(D+1)}$, $\boldsymbol{y} \in \mathbb{R}^N$. Assume that the intercept term is included in the $\boldsymbol{\theta}$ and that the weighted linear regression solution exists in this setting.

Questions:

1. (3 pts) Show that $J(\boldsymbol{\theta})$ can also be written as:

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

Here $\boldsymbol{W}$ is a diagonal matrix. What are the entries along the diagonal of $\boldsymbol{W}$?

$J(\theta)$ can be written in matrix form:

$$J(\theta) = \sum_{n=1}^{N} w_n (y_n - \theta^T x_n)^2$$

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} \in R^{N\times(D+1)}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad W = \begin{pmatrix} w \\ \vdots \\ w \end{pmatrix}$$

$$y - X(\theta) = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} x_1^T\theta \\ x_2^T(\theta) \\ \vdots \\ x_N^T(\theta) \end{pmatrix} = \begin{pmatrix} y_1 - X^T\theta \\ \vdots \\ y_n - X_N^T\theta \end{pmatrix}$$

Therefore multiply by matrix of weights w_n.

Now rewrite $J(\theta) = (y - X\theta)^T W (y - X(\theta))$

The diagonal entries of W should be the weights $w_{1\ldots n}$

Brandon L
10575

2. (4 pts) Show that the optimal value for $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}$.

We compact the expression :

$$J(\theta) = (y - X\theta)^T W (y - X\theta)$$

$$= (y^T - \theta^T X^T) W (y - X\theta)$$

$$= y^T W (y - X\theta) - \theta^T X^T W (y - X\theta)$$

$$= y^T y W - WX\theta - \theta^T X^T W y - \theta^T X^T W X\theta$$

$$= \text{constant} - 2y^T WX\theta + \theta^T X^T X\theta W$$

$$\nabla J(\theta) = 2X^T X \theta W + 2X^T y W$$

$$\hat{\theta} = (X^T X W)^{-1} W X^T y$$

Brandon L
10575?

3. (3 pts) In class, we provided a probabilistic interpretation of ordinary least squares (OLS). We now try to provide a probabilistic interpretation of weighted linear regression. Consider a model where each of the N samples is independently drawn according to a normal distribution

$$P(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y_n - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_n)^2}{2{\sigma_n}^2}\right)$$

In this model, each y_n is drawn from a normal distribution with mean $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_n$ and variance ${\sigma_n}^2$. The ${\sigma_n}^2$ are known. Write the log likelihood function (express your answer in terms of y_n, $\boldsymbol{x}_n$, $\boldsymbol{\theta}$, σ_n^2 and any constants).

$\mathcal{LL} = \log\left(P(y_n|x_n, \theta)\right)$

$= \log\left(\quad\right)$

Brandon Lo
10575356

4. (4 pts) Show that finding the maximum likelihood estimate of $\boldsymbol{\theta}$ leads to the same answer as solving a weighted linear regression. How do ${\sigma_n}^2$ relate to w_n?

MLE: Principle for finding good parameters by maximizing the likelihood of observing the given data

This would give the same output θ as weighted linear regression.

Brandon Lo
1057535

4 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

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

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

w^* is parallel to vector from $\phi(x_1)$ to $\phi(x_2)$, which is $\phi(x_2) - \phi(x_1)$

The optimal hyperplane should go in the middle of the two and be perpendicular to the line that connects them.

w^* should be parallel to $\phi(x_2) - \phi(x_1) = \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix}$

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

We find the midpoint then the distance from the midpoint to either sample.

$$\gamma = \frac{\| [1,0,0]^T - [1,2,2]^T \|_2}{2} = \frac{\sqrt{8}}{2} = \sqrt{2}$$

Brandon Le
1057535

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

We know that w^* takes the form of $\begin{pmatrix} 0 \\ x \\ x \end{pmatrix}$ for a scalar x and $Y = \sqrt{2}$

$$\text{margin} = \frac{1}{\|w^*\|_2} = Y = \sqrt{2}$$

$$\|w^*\|_2^2 = \left(\frac{1}{\sqrt{2}}\right)^2 = \frac{1}{2}$$

$$w^{*T} w^* = \frac{1}{2}$$

$$2x^2 = \frac{1}{2}$$

$$x = \frac{1}{2}$$

Therefore, $w^* = \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$

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

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

$$[0, \tfrac{1}{2}, \tfrac{1}{2}] \cdot [1, 0, 0]^T + b^* = -1$$

$$b^* = -1$$

$$f(x) = [0, \tfrac{1}{2}, \tfrac{1}{2}] \cdot [1, \sqrt{2}x, x^2]^T - 1$$

$$f(x) = \frac{x^2}{2} + \frac{\sqrt{2}x}{2} - 1$$

Brandon Lo
1057553

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

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

Brandon Lo
105753560

5 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 r_{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 centroid $\boldsymbol{\mu}_k$ can be updated as $\sum_{n=1}^{N} \alpha_{nk} \boldsymbol{x}_n$. Specifically, show how α_{nk} can be computed given all r_{nk}'s.

α_{nk} = mistakes make on example, kth cluster

Given all r_{nk}s, α_{nk} can be computed because r_{nk} shows which cluster it is assigned to, and thus how many mistakes there are

$$\alpha_{n,k} = \frac{r_{nk}}{\sum_{n=1}^{N} r_{nk}}$$

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.

$$d_{Kernel}(x_1, x_2) = k(x_1, x_1) + k(x_2, x_2) - 2k(x_1, x_2)$$

This is the distance between $\phi(x_1)$ and $\phi(x_2)$:

$$d_{Kernel}(x_1, x_2) = d(\phi(x_1), \phi(x_2))$$

$\phi(x_1)^T \phi(x_2)$ can give distance, which can be defined in terms of original features

Brandon L
105753

3. (4 pts) Given the results of the above two parts, show how to compute the squared distance between an arbitrary data point $\boldsymbol{x}_m$ and a centroid $\boldsymbol{\mu}_k$ ($\|\boldsymbol{x}_m - \boldsymbol{\mu}_k\|^2$) 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 pairs of data points).

$$\|x_n - \mu_k\|^2 = \|x_n - \sum_{n=1}^{N} \alpha_{nk} x_n\|^2$$

$$= x_n^T x_n + \sum_i \sum_j \alpha_{ik} \alpha_{jk} x_i^T x_j - 2 \sum_i \alpha_{ik} x_n^T x_i$$

$$\mu_k(\|x_n - \mu_k\|^2) = \frac{1}{N} \sum_{n=1}^{N} \alpha_{nk} x_n \cdot (\quad \downarrow \quad)$$

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

Brandon Lo
105753560

6 Naive Bayes (10 pts)

You are attempting to train a Naive Bayes model for spam classification of emails (with each email modeled as a bag-of-words). An email consists of features X_d (denoting word d in the document) where each word can take one of three values: a, b, or c. The label Y can take one of two values (1 = Spam or 0 = Ham).

1. (2 pts) What modeling assumption(s) does the Naive Bayes model make (Select all that apply)?

   (a) $P(X_1 = a, X_2 = b, X_3 = c) = P(X_1 = a)P(X_2 = b)P(X_3 = c)$.
   (b) $P(X_1 = a, X_2 = b, X_3 = c|Y = 1) = P(X_1 = a|Y = 1)P(X_2 = b|Y = 1)P(X_3 = c|Y = 1)$.
   (c) $P(X_1 = a, X_2 = b, X_3 = c|Y = 1) = P(X_1 = a, X_2 = c, X_3 = b|Y = 1)$
   (d) $P(X_1 = a, X_2 = b, X_3 = c, Y = 1) = P(Y = 1)P(X_1 = a)P(X_2 = b)P(X_3 = c)$

2. (2 pts) How many parameters does this Naive Bayes model contain (after accounting for constraints)?

   (a) 5
   (b) 6
   (c) 11
   (d) 18

Need to account for features a, b, or c and label is 1 or 0

Brandon Lo
105753561

You are given the following set of training examples to estimate the parameters:

Email	Y
a, c, a	1
c, a, c	1
a, a	0
b, c, b	0
c, c, b	0

3. (2 pts) What is the maximum likelihood estimate of $P(Y = 1)$?

   (a) $P(Y = 1) = \frac{1}{5}$

   (b) $P(Y = 1) = \frac{2}{5}$

   (c) $P(Y = 1) = \frac{3}{5}$

   (d) $P(Y = 1) = \frac{4}{5}$

4. (2 pts) What is the maximum likelihood estimate of the probability of a word taking the value "a" given $Y = 1$ ($P(a|Y = 1)$?

   (a) $\frac{1}{4}$

   (b) $\frac{1}{3}$

   (c) $\frac{2}{5}$

   (d) $\frac{1}{2}$

5. (2 pts) To classify a new document $\boldsymbol{x} = (a, a, a)$ using this model, we need to compute the posterior probability $P(Y = 1|\boldsymbol{x})$. What is the posterior probability $P(Y = 1|\boldsymbol{x})$?

   (a) $\frac{P(a|Y=1)P(Y=1)}{P(a|Y=1)P(Y=1)+P(a|Y=0)P(Y=0)}$

   (b) $\frac{P(a|Y=1)^3P(Y=1)}{P(a|Y=1)^3P(Y=1)+P(a|Y=0)^3P(Y=0)}$

   (c) $P(a|Y = 1)P(Y = 1)$

   (d) $P(a|Y = 1)^3P(Y = 1)$

Bayes rule:

$$\frac{p(x|Y=1)\,p(Y=1)}{p(x)}$$

Identities

Probability density/mass functions for some distributions

$$\text{Normal} \quad : \quad P(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\text{Multinomial} \quad : \quad P(\boldsymbol{x};\boldsymbol{\pi}) = \prod_{k=1}^{K} {\pi_k}^{x_k}$$

$\boldsymbol{x}$ is a length K vector with exactly one entry equal to 1 and all other entries equal to 0

$$\text{Poisson} \quad : \quad P(x;\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

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.