Temple University

DEPARTMENT OF COMPUTER INFORMATION SCIENCES

CIS4526: Foundations of Machine Learning Final Exam

SEMESTER 2019 Fall, Instructor: Kai Zhang

Time allowed: 80 minutes (closed book)

Student name ____________________ **Student ID** ________________________

Q1	Q2	Q3	Q4	Q5	Q6	Q7	All

1. Correlation.

a. Given two d-dimensional column vectors, x, y, how to compute their pearson correlation coefficient? (5 points)

Let $\tilde{x} = x - \bar{x}$, $\tilde{y} = y - \bar{y}$, where $\bar{x}, \bar{y}$ are the mean of vector x and y.

then correlation is $\frac{\tilde{x}^T \tilde{y}}{\sqrt{\|\tilde{x}\|_2 \cdot \|\tilde{y}\|_2}}$

b. Given a data matrix $X \in R^{n\times d}$, with n samples and d features, and label vector $Y \in R^{n\times 1}$ for regression. How to use correlation to select top-k useful features? (5 points)

Calculate the Correlation between Y and each feature in X, denoted by $X[:, i]$ (the ith column of X) as.

$corr_i$ = correlation$(X[:, i], Y)$.

Pick the top k features with highest $|corr_i|$.

2. Given a data matrix $X \in R^{n\times d}$, with n samples and d features.

a. How to calculate the covariance matrix, write down the formula clearly (5 points).

Define centering Matrix $C = I_{n\times n} - \frac{1 \cdot 1^T}{n}$.

Then $cov = (CX)^T (CX)$.

b. How to project all the samples to the dominant principal component? (3 points)

Eigenvalue decomposition $cov = U \Sigma U^T$. when U are the eigenvectors. Pick the one with largest eigenvalue, U_1, and the project as $X \cdot U_1$. $(U_1 \in R^{n\times 1})$

c. What are the main benefits of PCA in practical data analysis? (2 points)

denoising → improve subsequent classifier.

Improves efficiency.

3. Kernel trick is commonly used in nonlinear classification problems

a. Describe what is the kernel trick. (5 points)

Mapping x to $\Phi(x)$. by specifying the inner product of mapped data points with so called kernel function, i.e.

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle.$$

b. Suppose one adopts the degree-2 polynomial kernel $k(x,y) = (0.5 + x'y)^2$. Now, for two points in the original space $x = [1\ 0]'$ and $y = [-1\ 2]'$, suppose they have been mapped to the Hilbert space as $\Phi(x)$ and $\Phi(y)$ via the above polynomial kernel. Please compute the L2-norm distance of the two points in the Hilbert space $\|\Phi(x) - \Phi(y)\|_2$ (10 points)

$$= \langle \phi(x), \phi(x) \rangle + \langle \Phi(y), \Phi(y) \rangle - 2\langle \Phi(x), \Phi(y) \rangle.$$

$$= k(x,x) + k(y,y) - 2k(x,y)$$

$$= (0.5+1)^2 + (0.5+5)^2 - 2[0.5+(-1)]^2$$

$$= 2.25 + 30.25 - 0.5$$

$$= 32$$

so L_2-distance: is $\sqrt{32} = 4\sqrt{2}$

c. Write down the objective function of soft-margin SVM (in terms of constrained optimization) (10 points)

$$\min \frac{1}{\|W\|_2} + C\sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \begin{cases} (w^T x_i + b) y_i \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

4. Conditional probability and Information gain.

a. Write down Bayesian conditional probability P(X|Y) for random variables X and Y; (3 points)

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

b. Given the table between two random categorical variables, "Temperature" (X = High, Median, Low) and "Rain" (Y = Yes, No). Compute the following

Temperature(X)	Rain(Y)
High	Yes
High	Yes
Median	Yes
Median	Yes
Median	No
Median	No
Low	Yes
Low	No
Low	No
Low	No

- P(X = High) (2 points)

  0.2

- P(X=Low, Y = No) (2 points)

  0.3

- P(X = High|Y = Yes) (3 points)

  0.5

c. What is the entropy of the random variable Y? (3 points)

$-(0.5\log 0.5 + 0.5\log 0.5)$

≈ 1

d. What is the information gain on Y after partitioning it by variable X? (7 points)

X= High 20% | X= Median 40% | X=low 40%

entropy = 0 | entropy = 1 | entropy $= -(\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}) = 0.81$

Expected Entropy = $0 + 0.4\times 1 + 0.4\times 0.81 = 0.724$

$IG = 1 - 0.724 \approx 0.276$ bit

5. Logistic regression.

a. Explain what is maximum likelihood estimation, given the data D and your model/hypothesis h. (8 points)

$$h^* = \arg\max_h \; p(D|h)\cdot \underbrace{p(h)}_{\text{assumed uniform.}}$$

b. For a two-class problem, derive the maximum likelihood function for logistic regression. (7 points)

$$p(x_i \in +1 | h) = \frac{1}{1+e^{-f(x_i)}} \text{ (positive class)}$$

$$p(x_i \in -1 | h) = \frac{1}{1+e^{f(x_i)}} \text{ (negative class)}$$

The likelihood becomes $\prod_{x_i \in (+1)} \frac{1}{1+e^{-f(x_i)}} \cdot \prod_{x_j \in (-1)} \frac{1}{1+e^{f(x_j)}}$

Negative loglikelihood becomes $\sum_{i=1}^{N} \log\left(1+e^{-y_i f(x_i)}\right)$

6. Outlier detection. In unsupervised setting, outliers are those samples that are far away from others. Given a sample set $\{x_i\}$ for I = 1,2,..., n. If we assume that the samples are drawn from a Gaussian distribution, how to detect the outliers? (10 points)

The mean and ~~co~~variance matrix of the Gaussian distribution can be estimated as follows

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$$

Then the density is $f(x) = \frac{1}{\sqrt{2\pi}|\Sigma|} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$

plugging x_i's into the density funtion $f(x)$

those x_i's with minimum density values will be selected as outliers based on the threshold

7. K-means clustering is a popular unsupervised learning method.

a. What is the loss function of k-means? (5 points)

$$L = \sum_{i=1}^{N} W_{ik} \| x_i - u_k \|^2 .$$

if x_i belongs to cluster k, $W_{ik} = 1$, or else $W_{ik} = 0$

b. What is the E step and M step of the k-means algorithm? (5 points)

E. step. assign x_i to the closest cluster center u_k

M step. $$u_k \leftarrow \frac{\sum_{i=1}^{N} W_{ik} \cdot x_i}{\sum_{i=1}^{N} W_{ik}}$$

c. **(bonus)** Derive the M step (i.e., show that why we should perform M step that way). (10 points)

The loss function can be decomposed into K independent parts

$$L = \sum_{k=1}^{K} \sum_{x_i \in C_k} W_{ik} \| x_i - u_k \|^2$$

for one of the K terms. Compute the gradient.

$$L_k = \sum_{x_i \in C_k} W_{ik} \| x_i - u_k \|^2$$

$$= \sum_{i} W_{ik} (x_i - u_k)^T (x_i - u_k)$$

$$\frac{\partial L_k}{\partial u_k} = \sum_{i} W_{ik} (x_i - u_k) = 0 \quad \text{(gradient set to 0)}$$

$$\Downarrow$$

$$u_k = \frac{\sum_{i} W_{ik} . x_i}{\sum_{i} W_{ik}}$$