Student Name: Student Number:

1. (30 points) **TRUE/False questions.** Which of the following statements are correct and which are false? Clearly write "TRUE" or "FALSE" besides each of the following options.

True ○ Mapping a data set into a higher dimensional space (e.g., by using polynomial mappings) can help with the underfitting issue for a downstream supervised learning task (e.g., regression).

False ○ The decision boundary of nearest neighbor classifier is linear.

False ○ We map a data set from d-dimensional space to q dimensional space using Principal Component Analysis (PCA). Choosing a larger value of q will increase the risk of losing more information.

False ○ Linear programming can be used to find a linear classifier with minimum classification error even when the data points are not linearly separable.

True ○ The least-squares problem will have a unique solution when $\text{Rank}(X) = d$, where d is the dimension of each data point, N is the number of data points, and X is the N-by-d data matrix.

False ○ We are using a kernel that corresponds to the mapping $\phi(x) : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$. The kernel trick is especially useful when the number of data points is much larger than d_2.

True ○ k-nearest neighbor classifier is less likely to overfit compared with nearest neighbor classifier.

True ○ Consider the regularized least squares approach. Increasing the coefficient of the regularization penalty would decrease the chance of overfitting.

False ○ The perceptron algorithm will converge even if the data is not linearly separable.

False ○ We want to do regression and we know the exact distribution that the data is being generated from (i.e., we know $P(x, y)$). In this case, whether the goal is minimizing the squared loss (i.e., $(y - \hat{y})^2$) or the absolute loss (i.e., $|y - \hat{y}|$) would not affect the answer.

2. Consider the bias-variance-noise trade off that we discussed in the class. We showed that the expected squared error of any curve-fitting approach can be decomposed into three terms:

1. $E_{x,y}(y - y^*(x))^2$
2. $E_x(y^*(x) - E_Z(\hat{y_Z}(x)))^2$
3. $E_{x,Z}\left(\hat{y_Z}(x) - E_{Z'}(\hat{y_{Z'}}(x))\right)^2$

(a) (6 points) Which of these terms indicate the noise? Which one is the bias? Which one is the variance?

(1) noise

(2) bias

(3) Variance

(b) (3 points) We are comparing the Least Squares (LS) approach and the Regularized Least Squares (RLS) approach for curve-fitting. Which of the above three terms is surely identical across the two approaches? (i.e., takes the same value for LS and RLS)?

(1)

(c) (3 points) Which of the three terms is likely to increase if we use RLS with a very large regularization penalty? (compared with just using LS)

(2)

(d) (3 points) We are using the LS approach for two different curve-fitting problems (i.e., on two different distributions, $P_1(X,Y)$ and $P_2(X,Y)$). Assume the two marginal distributions are identical, i.e., $P_1(X) = P_2(X)$. However, the conditional distributions are different, i.e., $P_1(Y|X) \neq P_2(Y|X)$. Which of the above three terms would remain unchanged?

(3)

3. We are given a (training) data set $Z = \{(x_i, y_i)\}_{i=1}^n$ consisting of n data points, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We would like to "fit" a (homogeneous) linear curve ($\hat{y} = W^T x$) to the data, where $W \in \mathbb{R}^d$. For this, we want to use a probabilistic model.

   (a) (15 points) Assume $P(y|x, W) = \frac{1}{\sqrt{(2\pi)^d}} exp\left(-\frac{\|y - W^T x\|_2^2}{2}\right)$ and $P(x, W) = P(x)P(W)$. What is the maximum likelihood solution for W in terms of Z? What is another name for this approach of finding W? Show your work.

$$\arg\max P(Z|w) = \arg\max \prod_{i=1}^{n} P(x^i, y^i | w)$$

$$= \arg\max \frac{\prod_{i=1}^{n} P(y^i | x^i, w)\, P(x^i)\, \cancel{P(w)}}{\cancel{P(w)}}$$

$$= \arg\max_{w} \prod_{i=1}^{n} P(y^i | x^i, w)$$

$$= \arg\max_{w} \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{\|y^i - w^T x^i\|_2^2}{2}\right)$$

$$= \arg\max_{w} \sum_{i=1}^{n} \|w^T x^i - y^i\|_2^2$$ Typo: Put arg min instead of arg max

Taking derivative $\longrightarrow$ $w^* = (X^T X)^{-1} X^T Y$

LS

CamScanner

(b) (20 points) For this part, we still assume $P(y|x,W) = \frac{1}{\sqrt{(2\pi)^d}} exp\left(-\frac{\|y-W^T x\|_2^2}{2}\right)$ and $P(x,W) = P(x)P(W)$. On top of that, assume $P(W) = \frac{1}{2^d} exp\left(-\|W\|_1\right) = \frac{1}{2^d} exp\left(-\sum_{i=1}^{d} |w_i|\right)$. What is the maximum a posteriori solution for W in terms of Z? Show your work.

$$\arg\max_w P(w|Z) = \arg\max_w P(Z|w)P(w)$$

$$= \arg\max_w \left[\ln p(w) + \sum_{i=1}^{n} \ln p(x^i, y^i | w)\right]$$

$$= \arg\max_w \left[\ln P(w) + \sum_{i=1}^{n} \ln \frac{P(y^i|x^i,w)P(x^i)\cancel{P(w)}}{\cancel{P(w)}}\right]$$

$$= \arg\max_w \left[\ln P(w) + \sum_{i=1}^{n} \ln P(y^i|x^i, w)\right]$$

$$= \arg\max_w \left[d\ln\left(\tfrac{1}{2}\right) - \|w\|_1 - \frac{nd}{2}\ln(2\pi) - \sum_{i=1}^{n} \frac{\|y^i - w^T x^i\|_2^2}{2}\right]$$

$$= \arg\min_w \left(\|w\|_1 + \frac{1}{2}\sum \|y^i - w^T x^i\|_2^2\right)$$

$$= \arg\min_w \left(2\|w\|_1 + \sum \|y^i - w^T x^i\|_2^2\right)$$

4. (20 points) In this question we will use least squares to find the best line ($\hat{y} = ax$) (so $a \in \mathbb{R}$ is the parameter we are looking for) that fits the function $f(x) = x^2$. Assume x is uniformly distributed on $[1, 2]$. Find value of a that minimizes the expected squared error. Show your work.

$$a^* \underset{a}{\text{argmin}} \int_1^2 1 \times (x^2 - ax)^2 dx$$

$$\frac{\partial}{\partial a} = \frac{\partial}{\partial a} \int_1^2 (x^2 - ax)^2 dx = \int_1^2 2(x^2 - ax)(-x) = 0$$

$$\int_1^2 (x^3 - ax^2)\, dx = 0$$

$$\left[\frac{x^4}{4} - \frac{ax^3}{3} \right]_1^2 = 0$$

$$4 - \frac{8a}{3} - \left(\frac{1}{4} - \frac{a}{3} \right) = 0$$

$$\frac{15}{4} - \frac{7a}{3} = 0 \qquad a = \frac{3 \times 15}{4 \times 7} = \frac{45}{28}$$

$$\approx 1.607$$