1 SHORT QUESTIONS (30PT)

(a) (2pt) When we run K-Means on the same dataset with different starting conditions, we can get different clusterings. True or False?

Explanation: True, because K-Means can fall into local minimas

(b) (2pt) Linear Discriminant Analysis assumes that the class conditioned densities are Gaussian with the same covariance matrix. True or False?

Explanation: True, When the covariances in GDA are equal, we have an algorithm called Linear Discriminant Analysis or LDA.

(c) (2pt) Kernelized SVM and Kernelized Ridge Regression are examples of non-parametric models. True or False?

Explanation: False, KSVM and KRR are parametric models

(d) (2pt) KNN is an example of supervised learning method. True or False?

Explanation: True, KNN is a supervised learning method

(e) (2pt) Suppose we are running the SVM algorithm on a binary classification task that has labels +1 and -1 . Let f be the score function that maps inputs x to real numbers. Using the hinge loss, for a data point (x, y) that has label value $y = +1$, all predictions $f(x) \geq 1$ will incur the same loss. True or False?.

Explanation: True, SVM does not penalize correct predictions

(f) (2pt) List three methods commonly used for handling overfitting:

Explanation: Use a simpler model family, collect more training data or modify the training process to penalize overly complex models

(g) (2pt) Given a design matrix X, the normal equations $\left(X^\top X\right)^{-1} X^\top y$ always involve an invertible matrix $\left(X^\top X\right)$. True or False?

Explanation: False, we only assume the matrix is invertible

(h) (2pt) The kernel trick can only be used to efficiently compute the dot product between finite-dimensional features $\phi(x)$. True or False?

Explanation: False, We can efficiently use any features $\phi(x)$ up to infinite dimensions i.e. Taylor Series

(i) (2pt) A Kaggle competition typically has two different test sets. One set is used to score the public leaderboard during the competition. The other set is used to create a private leaderboard after the competition ends. One team repeatedly submits their method on the public set, making small changes each time until they achieve first place on the public leaderboard. When the competition ends, teams are re-scored on the private set. The team finds they have dropped to 102^{nd} place. What happened? Please describe.

Explanation: The team has overfitted their model that performs extremely well on a public set. However, since their data was trained exclusively on a public set, the private set did not fit their model well, rather than the opposite.

For each of the listed descriptions below, choose whether the experimental set up is ok or problematic. If you think it is problematic, briefly state what the problems are:

(j) (4pt) A team is building a 10-class digit classifier. They're trying out a new method of cross-validation. Instead of choosing the folds randomly, they constrain each fold to contain only a single digit class. Ok or Problematic?

Explanation: Problematic, By constraining each fold to contain only a single digit class, the training set for each fold will lack 9 out of the 10 classes

(k) (4pt) A team is using k-nearest neighbor to classify birds. They use 10 -fold cross validation to pick the k that maximizes the average validation accuracy. Ok or Problematic?

Explanation: Ok, Using k-fold cross-validation to pick the k that maximizes the average validation accuracy is a standard and widely accepted approach for hyperparameter tuning.

(l) (4pt) Suppose an intern at Google is trying to build a better spam classifier. They believe their model is good because every spam message in their test set was caught and correctly classified as spam. Ok or Problematic?

Explanation: Problematic, if their model catches every single spam message then they are likely overfitting their model to cater to the test set

2 Probability and MLE [20 POINTS]:

Let x be the time interval (in hour) between two messages on CS5785 Slack team. Assume that x has the following probability density function:

$$p(x;\theta) = \begin{cases} \theta^{-1}e^{-x/\theta}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

You check the Slack team several times and get 4 independent observations of $x : x_1 = 0.2, x_2 = 0.3, x_3 = 0.5, x_4 = 1$.

(a) (10pt) What is the log-likelihood of your observations given $\theta = 1$?

Answer Plug in your observations $x_1 = 0.2, x_2 = 0.3, x_3 = 0.5, x_4 = 1$ and $\theta = 1$ to get the likelihood:

$$L(1; 0.2, 0.3, 0.5, 1) = p(0.2; 1) \times p(0.3; 1) \times p(0.5; 1) \times p(1; 1)$$
$$= e^{-0.2} \times e^{-0.3} \times e^{-0.5} \times e^{-1}$$

Take the natural log to get the log-likelihood:

$$l(1; 0.2, 0.3, 0.5, 1) = \ln(e^{-0.2} \times e^{-0.3} \times e^{-0.5} \times e^{-1})$$

$$= (-0.2) + (-0.3) + (-0.5) + (-1)$$

$$= -2$$

So, the log-likelihood of the observations given $\theta = 1$ is -2.

(b) (10pt) What value of θ maximizes the likelihood? Provide justification via mathematical deduction.

Answer To find the value of θ that maximizes the likelihood, we need to take the derivative of the log-likelihood with respect to θ, set it to zero, and solve for θ. This will give us the maximum likelihood estimate (MLE) of θ.

The log-likelihood is:

$$l(\theta) = \sum_{i=1}^{4} \ln(p(x_i; \theta))$$

$$= \sum_{i=1}^{4} \left(-\ln(\theta) - \frac{x_i}{\theta} \right)$$

$$= -4\ln(\theta) - \frac{1}{\theta} \sum_{i=1}^{4} x_i$$

Differentiating with respect to θ:

$$\frac{\partial l(\theta)}{\partial \theta} = -\frac{4}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{4} x_i$$

Setting the derivative to zero to find the extrema:

$$-\frac{4}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{4} x_i = 0$$

$$-4\theta + \sum_{i=1}^{4} x_i = 0$$

$$\implies \theta = \frac{\sum_{i=1}^{4} x_i}{4}$$

$$\theta = \frac{0.2 + 0.3 + 0.5 + 1}{4} = \frac{2}{4} = 0.5$$

3 DECISION BOUNDARIES (15PT)

a (2pt each; 10 total) Here are several binary classification algorithms. Match the decision boundary with the classification method; you should write down the letter cor-

responding to the plot next to each classifier name. Each plot corresponds to one algorithm on the left; select the best match.

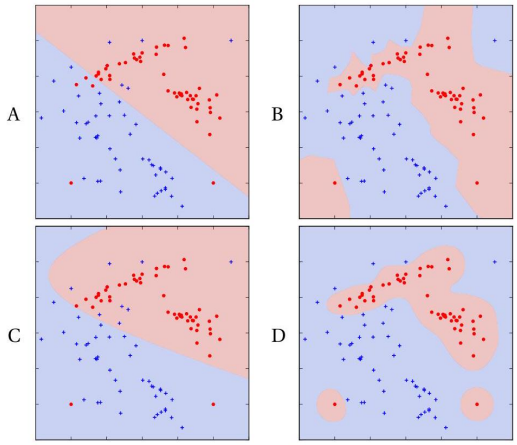

Figure 1: Enter Caption

- k-NN, for $k = 1$:

Answer: B, This classifier would have boundaries that are extremely flexible and sensitive to individual data points, effectively "hugging" the data closely.

- SVM, polynomial kernel kernel:

Answer: C, The SVM with a polynomial kernel would produce curved decision boundaries, but they would typically be smoother and more systematic than the extremely flexible boundaries of a 1-NN classifier.

- Linear discriminant analysis:

Answer: A, LDA tries to find a linear combination of features that characterizes or separates two or more classes. Therefore, the decision boundary is linear.

- SVM, radial basis function kernel:

Answer: D, The RBF kernel in SVM can create non-linear decision boundaries and is capable of forming more complex shapes, like circles or ellipses around data clusters.

b (5pt) Which classifier has the most similar decision boundary to an SVM with the primal formulation?

Answer The SVM with the primal formulation refers to an SVM that uses a linear kernel. The primal problem for SVMs represents a linear decision boundary in the input space.

Given this, the classifier with the most similar decision boundary to an SVM with the primal formulation is the one with a linear decision boundary.

From the provided plots, Plot A shows a clear linear boundary. Therefore, the classifier that has the most similar decision boundary to an SVM with the primal formulation is the one represented by Plot A, which is the Linear discriminant analysis (LDA).

4 Principal Component Analysis (15 PT)

Given four 2D data points $(-2,-2),(2,2),(-1,1),(1,-1)$, and we want to perform Principal Component Analysis (PCA) to project the data points into compact 1D space.

1. (5pt) Find the first principal component of the four 2D data points. This is a 2D eigenvector with the largest eigenvalue in eigendecomposition. **Hint:** *You can first plot the data points and think about the alignment of data points visually. It's possible to obtain the principal components without manually performing eigendecomposition so you should not need to do that. Also, if a vector is the first principal component, its negation is also the first principal component, so feel free to provide just one of them.*

Answer: Vector of y=x. [1, 1]

2. (5pt) The objective of PCA is to choose a projection direction that maximizes the variance of projected data points. What is the variance of the projected 1D data points?

Answer: Compute the variance:

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Where x_i are the data points and μ is the mean of the data points.

For our data, the mean, μ, is 0 (since the mean of the projected coordinates is 0).

Thus, the variance is:

$$\text{Variance} = \frac{1}{4}[(-2\sqrt{2})^2 + (2\sqrt{2})^2 + 0^2 + 0^2]$$

$$\text{Variance} = \frac{1}{4}[8 + 8] = 4$$

Therefore, the variance of the projected 1D data points is 4.

3. (5pt) An alternative objective of PCA is minimizing the reconstruction error. What is the reconstruction error? **Hint:** *Recall that given N data points $x^{(i)} \in \mathbb{R}^2$ (for $i = 1,\ldots,N$) and the orthonormal projection matrix $W \in \mathbb{R}^{2\times 1}$, the reconstruction error is given by $J(W) = \sum_{i=1}^{N} \left\|x^{(i)} - WW^T x^{(i)}\right\|_2^2$.*

Answer: The reconstruction error for points 1 and 2 is 0 since they lie on the vector line. For points 3 and 4, it is 2 since the squares of the distance between the vector and the point is 2. Thus the total is $0 + 0 + 2 + 2 = 4$

5 NAÏVE BAYES WITH BAG-OF-WORDS (20PT)

We are building a twitter robot that can tell the difference between tweets about sports versus academic tweets about machine learning conferences. Our students have scoured twitter, collecting a dataset as follows. Each tweet is labeled with either sports or machine learning.

Sports	*Machine learning*
Red Sox **win** big last week, **score** 36-35 against Yankees	**Deep Learning** models using rectified linear **loss** functions **score** better than hand-selected baselines in latest competition!
Learning their place on the new food chain, patriots suffer another **deep loss**	More **deep** model magic! Yann LeCun for the **win**!
Giants **score** 42 points for the **win**, averting another **loss** in the semifinals	Reinforcement **learning** is the wave of the future

a (10pt) Convert the above tweets into a bag of words representation, a binary vector with five elements. Calculate P (word i present|class), for each of the important bolded words: **Win**, **Loss**, **Learning**, **Score**, **Deep**.

Answer:

b (10pt) Consider the following tweet:

Latest ResNet is crushing the ImageNet competition, with lower **loss** value than other models! Another big **win** for **deep learning**!

Let x be its bag-of-words vector. What probability $P(\text{class}_{\text{ML}} \mid x)$ and $P\left(\text{class}_{\text{sparts}} \mid x\right)$ would a Naive Bayes classifier assign to this tweet? What label would it assign?

Answer: