I. Math and Probability Basics

Q1: Review of Definitions

(a) For a scalar random variable x, give the definition of its mean and variance.

(b) For a vector random variable $x \in \mathbb{R}^n$, give the definition of its mean and covariance.

Q2: Short questions

(a) Given a vector $v \in \mathbb{R}^2$ and a line orthogonal to the vector $u \in \mathbb{R}^2$, draw the orthogonal distance from v to the line, and write down the equation for the distance.

(b) A *hyperplane* is a subspace whose dimension is one less than that of its ambient space. Suppose the line in (c) was instead an n-1 dimensional hyperplane in the ambient space $\mathbb{R}^n$. What is the orthogonal distance from a point $u \in \mathbb{R}^n$ to the hyperplane?

(c) Show that for a matrix A and vector x, $\frac{\partial}{\partial x}(A^{-1}) = -A^{-1}\left(\frac{\partial}{\partial x}A\right)A^{-1}$. Use the fact that for any two matrices A and B, $\frac{\partial}{\partial x}(AB) = \frac{\partial A}{\partial x}B + A\frac{\partial B}{\partial x}$.

II. General Machine Learning Concepts

Q1. True or False questions

Circle the correct answer (T or F) and give a one-sentence explanation of each answer; answers without explanation will not be given full points.

a) An advantage of normal equations to solve linear regression is that choosing a step size (learning rate) is not necessary [T/F]

b) Maximum likelihood can be used to derive a closed-form solution to logistic regression [T/F]

c) The gradient descent update for logistic regression is identical to linear regression [T/F]

d) Changing the prior in Linear Discriminant Analysis changes the direction of the decision hyperplane [T/F]

Q2. Short answer questions

Answer the following questions in brief one to two sentence answers.

a) For a training dataset $D = \{x_i, y_i\}$ where x_i are the inputs and y_i are the output, explain the difference between discriminative and generative classification models.

b) Give one example of a discriminative classification model.

c) Give one example of a generative model.

d) Suppose you want to use training data D to adjust the parameters w of a model where $L(D) = p(D; w)$ is the likelihood of the data. You want to prevent overfitting using a squared norm regularizer. What should your objective function look like? Should you minimize or maximize it?

e) What is cross-validation?

f) How can we use cross-validation to prevent overfitting? Explain the procedure using the setup of (d).

Q3. Error metrics

   a) Give one example each of error metrics that can be used to evaluate: classification, regression, clustering, and reinforcement learning.

   b) Which are the correct definitions of precision and recall? Here 'actual positives' are examples labeled positive (by humans), and 'predicted positives' are examples for which the algorithm predicts a positive label.

1. $precision = \dfrac{true\ positives}{predicted\ positives}$ 2. $precision = \dfrac{true\ positives}{actual\ positives}$

3. $recall = \dfrac{predicted\ positives}{actual\ positives}$ 4. $recall = \dfrac{true\ positives}{actual\ positives}$

   c) Suppose your data has binary labels, where the negative class (y=0) occurs 99% of the time and the positive class (y=1) occurs 1% of the time. Which of the following results obtained by your machine learning algorithm *by itself* confirms that it is performing well on this data (circle one)? *Hint: think about what happens if your algorithm always outputs '0', or always outputs '1'.*

1. High Recall   2. High Accuracy   3. High F1 score

III. Bayesian Methods

Q1. Bias-variance in Bayesian models

Alice has a dataset of m points with n-dimensional inputs and scalar outputs. She has trained several regularized linear regression models using regularization parameters $\lambda = e^0, e^{-1}, e^{-2}, e^{-3}$.

   a) Which parameter will lead to highest bias? To highest variance?

   b) Alice then decides to use a Bayesian approach to control the complexity of the model. What is the Bayesian equivalent to changing λ?

c) Which Bayesian model should she use? Explain what makes the model Bayesian.

d) Alice fit the parameters of her model and wants to use the predictive distribution on new inputs. Explain what the predictive distribution is.

e) What is the difference between the predictive distribution of the ML-based and the Bayesian linear regression models?

Q2. Bayesian Maximum A Posteriori (MAP)

Suppose we are estimating the probability of seeing 'heads' (x = 1) or 'tails' (x = 0) after tossing a coin, with μ being the probability of seeing 'heads'. The probability distribution of a single binary variable $x \in \{0,1\}$ that takes value 1 with probability μ is given by the *Bernoulli* distribution

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$$

Suppose we have a dataset of independent coin flips $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ and we would like to estimate μ using the Bayesian MAP solution. Recall that we can write down the data likelihood as

$$p(x^{(i)}|\mu) = \mu^{x^{(i)}}(1-\mu)^{1-x^{(i)}}$$

Consider the following Bayesian prior on μ, which believes that the coin is either fair, or slightly biased towards 'tails':

$$p(\mu) = \begin{cases} 0.5 \ if \ \mu = 0.5 \\ 0.5 \ if \ \mu = 0.4 \\ 0 \ otherwise \end{cases}$$

Write down the MAP estimate for μ under this prior as a function of the likelihood and the prior. *(Hint: use the argmax function).*

IV. Unsupervised Learning

Q1. Short questions

a) For which problems below would anomaly detection be a suitable algorithm? Explain.
   1. Given an image of a face, determine whether or not it is the face of a particular famous individual.
   2. From a large set of hospital patient records, predict which patients have a particular disease (eg. flu)
   3. Given a dataset of credit card transactions, identify transactions to flag them as possibly fraudulent.
   4. From a large set of primary care patient records, identify individuals who might have unusual health conditions.

b) Suppose you have trained an anomaly detection system for intruder detection in a security camera, and your system flags anomalies when p(x) is less than ε. You find on the cross-validation set that it is missing many intruder events. What should you do?

c) Which of the following approaches use density estimation? Explain
   1. Linear discriminant analysis
   2. Logistic Regression
   3. Anomaly detection
   4. Generative Adversarial Networks

Q2. Generative adversarial network (GAN) mini-max loss

A GAN consists of a generator and a discriminator: the generator network $G(z)$ has parameters θ_G, takes a random noise vector z as input, and outputs a sample of data x (e.g., an image); the discriminator network $D(x)$ has parameters θ_D, takes in x, and outputs a binary label $y \in \{0,1\}$ where 1 indicates that the input is real and 0 that it is fake. The discriminator's cost function is the cross-entropy loss for the binary classification task over the real and the generated inputs:

$$J_D(\theta_D) = \mathbb{E}_x[-\log D(x)] + \mathbb{E}_z[-\log(1 - D(G(z)))]$$

Suppose the generator's loss is the opposite (negative) of the discriminator's cross entropy loss.

a) The generator's loss is shown in the plot above. Write down the equation for it.
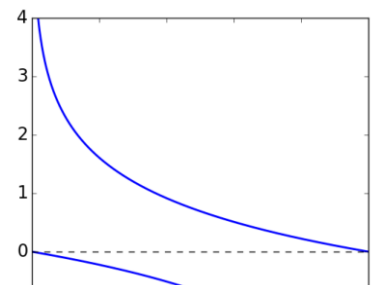
b) Explain why the generator's loss J_G suffers from slow learning.

c) Why doesn't this happen in the regular cross entropy loss?

Q4. [20 points] Neural Network for Binary Addition

(a) [5 points] Construct a simple neural network that computes the NOT operation of a single binary input variable x1 (0 or 1). Draw a diagram and indicate the values of the weights. Use the sigmoid activation function (shown in figure on the right), which ranges in [0, 1].
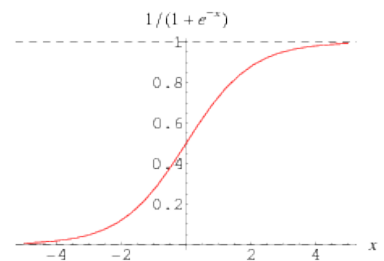
(b) [15 points] Design a neural network that adds two binary digits. The inputs are two single-digit binary variables, x1 and x2, and the outputs are the two digits y1, y2 of the sum of x1 and x2, i.e. x1+x2=s, where s has two digits, y1 and y2. The table below shows the correct outputs for each pair of inputs:

input		output	
x1	x2	y1	y2
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

i.e. y2 is the sum and y1 is the "carry." We have given sigmoid networks corresponding to the AND and OR functions above. Note that (A XOR B) can be expressed as (A OR B) AND NOT(A AND B)). Draw the complete addition network with all weights, indicating clearly the nodes for x1, x2, y1, y2. Be careful to include the bias unit(s).

Q4. Gradient Descent

For general hypothesis, θ, and cost $J(\theta)$, say whether each statement regarding gradient descent below is TRUE or FALSE, and give a one-sentence explanation.

a) [3 points] The cost function $J(\theta)$ is guaranteed to decrease with every iteration, regardless of the step size α. Circle one: TRUE / FALSE

b) [3 points] Convergence can be determined by looking at the change in the cost function across iterations. Circle one: TRUE / FALSE

c) [3 points] If gradient descent is converging very slowly, a smaller step size α should be used. Circle one: TRUE / FALSE

d) [3 points] Convergence of gradient descent to a global minimum is always guaranteed for convex cost functions. Circle one: TRUE / FALSE

V. Loss Functions, SVM, Kernels

Q1. Hinge Loss

Given the SVM decision function $f(x) = w^T x + b$ and labels $y \in \{-1, +1\}$, which of the following four loss functions implements the SVM hinge loss? Explain in words what the hinge loss does.

a.   b.

d.

https://powcoder.com
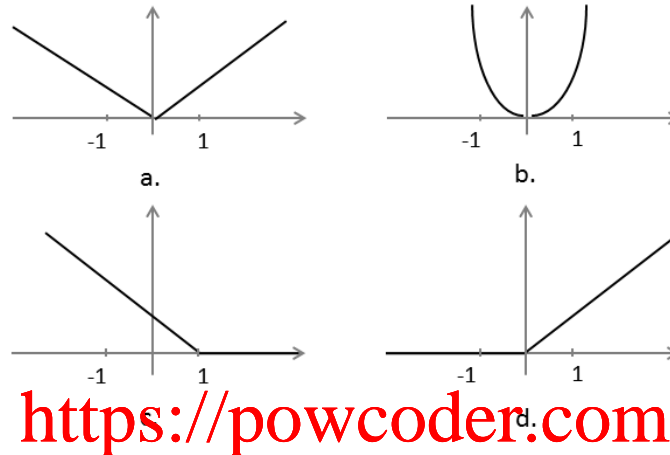
Assignment Project Exam Help

Q2. Loss functions for car sales

Alice works for a used car dealership. Her boss wants her to estimate the price y to charge for a car (in dollars) based on features such as: x_1 = car manufacturer, x_2 = model, x_3 = distance driven in miles, x_4 = age in years, etc. She collects data points from previous car sales $(x^{(i)}, y^{(i)})$, $i = 1, \dots, m$, where $y^{(i)}$ is the price the car was sold for, and decides to use a linear regression model, $y = \sum_{j=0}^{n} \theta_j x_j$.

a) Alice decides to minimize a sum of squares loss function. Does it make sense in this case? What effect will it have?

b) Her boss tells Alice she strongly prefers that the dealership not lose money on a sale. Given that the ith car cost the dealer $z^{(i)}$, suggest a way to pre-process the training data to encourage this.

Q3. Kernels and PCA

Consider the following dataset of 1-dimensional datapoints:

a) Which placement of Gaussian basis functions corresponds to a kernel feature representation for this

a.   b.

c.   d.
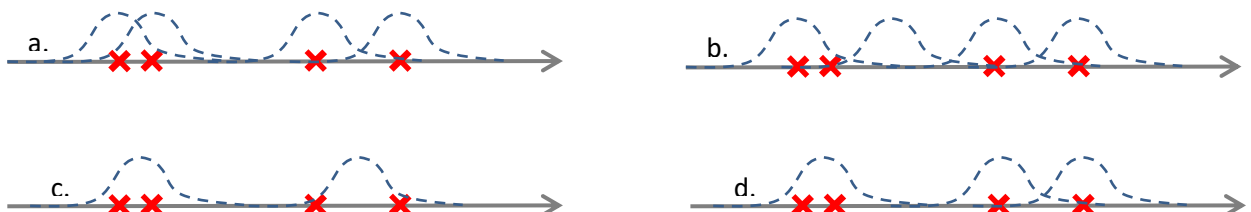
dataset?

For $k(x_1, x_2)$ to be a valid kernel, there must be a feature basis function $\varphi(.)$ such that we can write $k(x_1, x_2) = \varphi(x_1)^T \varphi(x_2)$.

b) Prove that $k = x_1^T C x_2$ is a valid kernel, where $C = XX^T$ is the data covariance obtained from the design matrix X. *Hint: use $(AB)^T = B^T A$*

c) Typically we use kernels to project data into a higher dimensional space. Principal Component Analysis projects data into a lower dimensional space. Suggest a kernel that uses PCA to project data into a lower, k-dimensional space, and prove that it is a valid kernel. You can assume the data already has a mean of zero. *Hint: use the result of (b).*
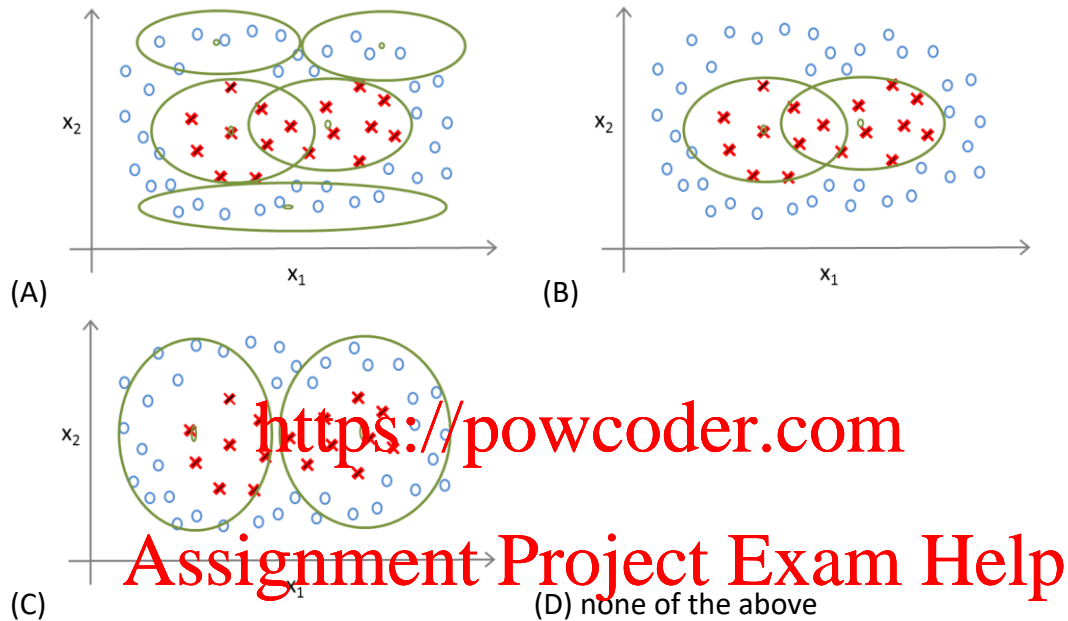
Q4. SVMs and Kernels

Consider the following dataset of 2-dimensional datapoints:
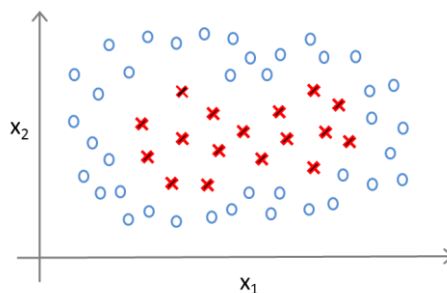
a) Which placement of Gaussian basis functions corresponds to a kernel feature representation for this dataset? Explain your answer in one sentence below.

(A)

(B)

(C)

(D) none of the above

b) How does increasing the variance of the Gaussian σ^2 affect the bias and variance of the resulting Gaussian Kernel SVM classifier? Explain.

c) For $k(x_1, x_2)$ to be a valid kernel, there must be a feature basis function $\varphi(.)$ such that we can write $k(x_1, x_2) = \varphi(x_1)^T \varphi(x_2)$. Suppose $k_1(x_1, x_2)$ and $k_2(x_1, x_2)$ are valid kernels. Prove that the following is also a valid kernel:

$$k(x_1, x_2) = k_1(x_1, x_2) + k_2(x_1, x_2)$$

d) Both SVMs with Gaussian kernels and Neural Networks with at least one hidden layer can be used to learn non-linear decision boundaries, such as the boundary between positive and negative examples in the dataset above. Describe the main similarity and the main difference in how these two approaches achieve this.

e) Explain what slack variables are used for when training SVMs.

VI. Reinforcement Learning

Q1. Markov Decision Process

Consider an agent that learns to navigate a simple grid world, shown below. Suppose the Markov Decision Process is given as follows:

States: locations (x,y) on the map
Actions: move right →, left ←, up↑, down↓ or the "do nothing" action ○
Reward: +1 in state (4,4), -0.01 in all other states
Transitions: $P_{state,action}(state')$ is given as

policy π

(1,1)	(2,1)	(3,1)	(4,1)
(1,2)	(2,2)	(3,2)	(4,2)
(1,3)	(2,3)	(3,3)	(4,3)
(1,4)	(2,4)	(3,4)	(4,4)

→	→	→	↓
→	→	→	↓
→	→	→	↓
→	→	→	○

$$P_{(x,y),\rightarrow}(s') = \begin{cases} 0.9 \ if \ s' = (x+1,y) \\ 0.1 \ if \ s' = (x,y) \\ 0 \ otherwise \end{cases}$$

$$P_{(x,y),\leftarrow}(s') = \begin{cases} 0.9 \ if \ s' = (x-1,y) \\ 0.1 \ if \ s' = (x,y) \\ 0 \ otherwise \end{cases} \quad P_{(x,y),\uparrow}(s') = \begin{cases} 0.9 \ if \ s' = (x,y-1) \\ 0.1 \ if \ s' = (x,y) \\ 0 \ otherwise \end{cases} \quad P_{(x,y),\downarrow}(s') = \begin{cases} 0.9 \ if \ s' = (x,y+1) \\ 0.1 \ if \ s' = (x,y) \\ 0 \ otherwise \end{cases}$$

(a) Consider the policy π shown above (right). If the agent starts in state (1,1) and takes actions according to this policy, what is the probability that it will end up in state (4,4) after 6 steps? Explain your answer.

(b) For the same policy π, suppose the agent transitions into state (4,4) for the first time on step 9. What is the total reward collected by the agent? Assume the agent collects reward in initial state and there is no discounting of rewards.

(c) Is there a more optimal policy (i.e. with strictly higher reward) for the agent than π? If so, what is it?

(d) What is the main difference between supervised learning and reinforcement learning?

Appendix: Useful Formulas

Matrix Derivatives

For vectors x, y and matrix A,

$y = Ax$, then $\frac{\partial y}{\partial x} = A$

If $z = x^T A x$, then $\frac{\partial z}{\partial x} = x^T(A + A^T)$. For the special case of a symmetric matrix A, $\frac{\partial z}{\partial x} = 2x^T A$.

Single Dimension Normal Distribution

$$N(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Multivariate Normal Distribution

The p-dimensional multivariate normal distribution is given by

$$N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$