Question 1. Short Answers [42 MARKS]

Part (a) [2 MARKS]

Suppose you have a choice between two neural net architectures for a classification task. In the first layer, Architecture A has a convolution layer with M input units and N output units. Architecture B has a fully connected layer with M input units and N output units. Both architectures use the same activation function, and subsequent layers are identical between the two architectures.

Which architecture has more parameters? Why?

SOLUTION B has more parameters. Convolutional layers use weight sharing and local connections.

Part (b) [2 MARKS]

Consider logistic regression with L2 regularization. The weight for the L2 regularizer (λ) is a hyperparameter for this model. Suppose that we are optimizing the weights using gradient descent. Is it a good idea to choose a value for λ by minimizing accuracy on a validation set?

Circle the correct answer: Yes or No

Explain briefly:

SOLUTION

Yes.

If the model overfits, it will assign lower accuracy to test data.

Part (c) [2 MARKS]

What is *underfitting* in Machine Learning and how do we prevent it?

It is when model is too simple, does not fit the data. Both training and validation errors are high. To prevent, we should use a more complex model.

Part (d) [5 MARKS]

What is *overfitting* in Machine Learning? List four methods to prevent overfitting and explain when each method should be preferably used.

It is when model is too complex, fits perfectly to the training data (including noise) and training error is very low. However, the validation error is high (doesn't generalize well).

To prevent:

1. use more training data (preferable if more data is feasible)
2. use simpler model (not a good choice unless there is no other choice)
3. regularization (good method if more data is not available)
4. early stopping (good method if more data is not available and if regularization is not working well)

Part (e) [2 MARKS]

Why is it called *Logistic Regression* if it is a classification model? What are the outputs of a logistic regression model?

Because we are predicting the probability of belonging to class using logistic function (the output is a continuous number). The outputs are the probabilities of each class.

Part (f) [4 MARKS]

Explain the *Gradient Descent* algorithm. Write down the update formula in each iteration. Visualize and explain the effects of high and low learning rates.

Many optimization problems dont have a direct solution and we use an iterative algorithm to minimize cost function called Gradient Decent. We initialize the weights to something reasonable (e.g. random or all zeros) and repeatedly adjust them in the direction of steepest descent.

update formula:

$$w_j \leftarrow w_j - \alpha \frac{\partial \mathcal{J}}{\partial w_j}$$

High learning rate can cause divergence. Low learning rate can make learning very slow.

Part (g) [4 MARKS]

What are the most common activation functions for the hidden and output layers of a Neural Network. For the output layer, name the activation functions used for each model type (i.e. regression, multi-class classification, and multi-label classification). Write a short sentence for the reason behind using each function.

logistic/sigmoid function, Relu and Relu family, Softmax function. The outputactivation functions:

- hidden layer: Relu
- regression output : None
- multi-class classification output: softmax function or logistic function for binary classification
- multi-label classification output: logistic function

Part (h) [2 MARKS]

Explain *Auto-Differentiation* and its relationship with the Backpropagation algorithm.

Autodifferentiation performs backprop in a completely mechanical and automatic way. Many autodiff libraries: PyTorch, Tensorflow, Jax, etc

Part (i) [2 MARKS]

What are the benefits of *Convolutional Networks* versus *Fully Connected Networks* when we are dealing with images as the model input? List two benefits.

1. lower number of parameters
2. preserving spatial properties of the image

Part (j) [4 MARKS]

Explain *Equivariance* and *Invariance* properties of Convolutional Networks and their relationship with each layer type of the network (i.e. convolution layer and pooling layer).

- Equivariant: if you translate the inputs, the outputs are translated by the same amount. convolution layers make it equivariant.
- Invariant: if you translate the inputs, the prediction should not change. Pooling layers provide invariance to small translations.

Part (k) [3 MARKS]

Explain the difference between *Discriminative* and *Generative* classifiers. Mathematically explain how the classification task is done in a generative classifier.

- Discriminative approach: estimate parameters of decision boundary/class separator directly from labeled examples.
- Generative approach: model the distribution of inputs characteristic of the class (Bayes classifier).

for generative classifier we model $p(x|t)$, then apply bays rule to get $p(t|x)$ for each class, then we chose the class with maximum posterior probability

Part (l) [2 MARKS]

What is the Naive assumption in the *Naive Bayes Models*. At what stage will you use Bayes rule in this algorithm?

Nave assumption: the features are conditionally independent given the class. $p(x_1 \dots x_D|c) = p(x_1|c) \dots p(x_D|c)$
We use bays rule at inference / prediction time.

Part (m) [3 MARKS]

Explain *Maximum Likelihood Estimation* and *Maximum A-Posteriori (MAP) Estimation* methods. What assumption do we make for MLE that makes it different from MAP.

- MLE: The maximum likelihood criterion says that we should pick the parameters that maximize the likelihood of seen data.
- MAP: Finds the most likely parameters to maximize posterior probability give the seen data.

in MLE we assume the prior distribution is uniform.

Part (n) [2 MARKS]

Explain the *Local Minima* issue in the *K-Means* clustering algorithm. Visualize the problem with a simple example.

the objective function of K-means is non-convex. The coordinate descent algorithm on the objective function is not guaranteed to converge to the global minimum. Nothing prevents k-means getting stuck at local minima. We could try many random starting points.

Part (o) [3 MARKS]

Explain, in words, the *Exploration-exploitation tradeoff* in Reinforcement Learning. Provide one example of Exploration-Exploitation in the real world.

if the agent strictly follow the policy (exploitation), it may never find some good actions and never experience all states and actions. The agent exploits its incomplete knowledge of the world by chooses the best action (i.e., corresponding to the highest action-value), but occasionally (probability eps) it explores other actions.

One real-world example is restaurant selection. Exploitation is to go to favourite restaurant. Exploration is to try a new restaurant.

They can use any other examples.

Question 2. Sensitivity and Specificity [8 MARKS]

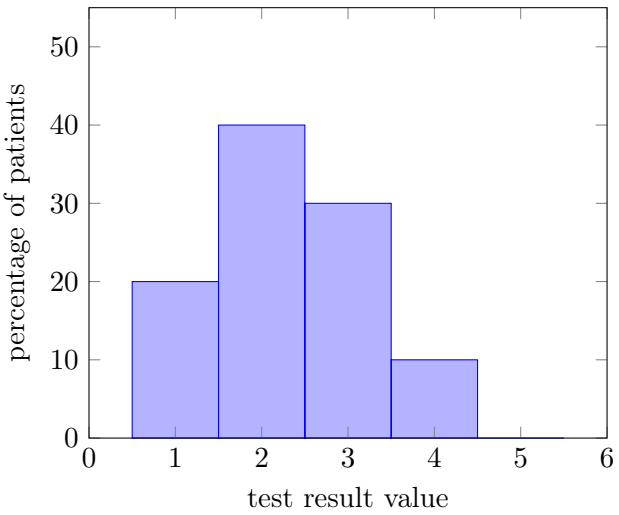
We want to build a model to diagnose a disease. Whether a patient has the disease is correlated with the result of Test A.

- Figure 1a shows the distribution of test result values given that the patient **does not have the disease**.
- Figure 1b shows the distribution of test result values given that the patient **has the disease**.

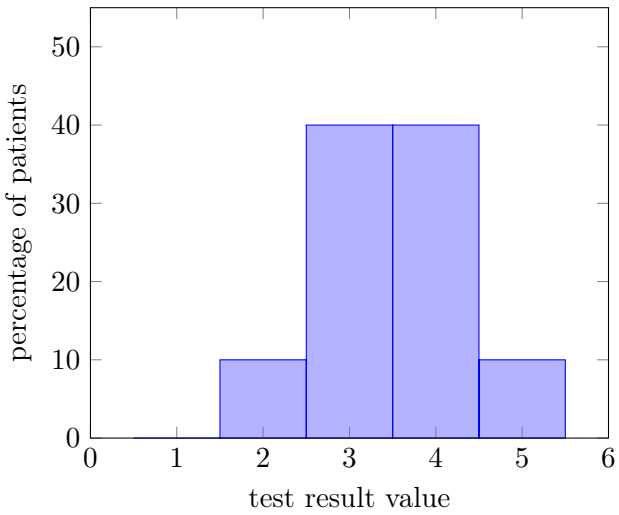
Our model makes the prediction based on a threshold value for the test.

For example, if we choose the threshold value to be 2.5, then the predictive rules are follows:

- If the patient's test result value is **greater than or equal to** 2.5,
  then our model predicts that the patient **has the disease**.
- Otherwise, if the patient's test result value is **less than** 2.5,
  then our model predicts that the patient **does not have the disease**.



(a) Test result if patient doesn't have disease D



(b) Test result if the patient has disease D

Figure 1: Test result distributions

Complete the table below. For each test threshold value, calculate the sensitivity and specificity of our predictive model. Recall that sensitivity is the true positive rate and specificity is the true negative rate. Express every value as a percentage.

Test Threshold Value	1.5	2.5	3.5	4.5
Sensitivity				
Specificity				

Table 1: Sensitivity and Specificity for Different Test Threshold Values

SOLUTION

The answers are in Table 2.

Test Threshold Value	1.5	2.5	3.5	4.5
Sensitivity	100%	90%	50%	10%
Specificity	20%	60%	90%	100%

Table 2: Solutions: Sensitivity and Specificity for Different Test Threshold Values

Question 3. Linear Regression [10 MARKS]

Consider a data set with three examples in Figure 2.

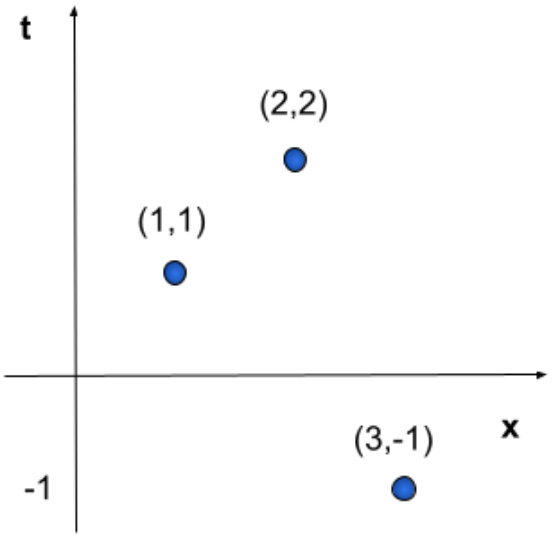


Figure 2: Data set for linear regression

We will fit a linear regression model to this data set. The closed-form solution for linear regression is given below.

$$w = (X^T X)^{-1} X^T t$$

Calculate the values of w for the linear regression model by using the formula above. We will give partial marks for correct intermediate steps.

Below is the formula to invert a 2x2 matrix.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

SOLUTION:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, t = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$$

$$X^T t = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{3 * 14 - 6 * 6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \begin{bmatrix} 7/3 & -1 \\ -1 & 1/2 \end{bmatrix}$$

$$(X^T X)^{-1} X^T t = \begin{bmatrix} 7/3 & -1 \\ -1 & 1/2 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 5/3 \\ -3/2 \end{bmatrix}$$

Thus, $w = -3/2$ and $b = 5/3$.

Question 4. Neural Networks [6 MARKS]

Part (a) [2 MARKS]

Your friend is training a feed-forward neural network model for a regression problem. They produced the following curve of training loss over iterations. Did your friend use stochastic gradient descent or batch gradient descent?

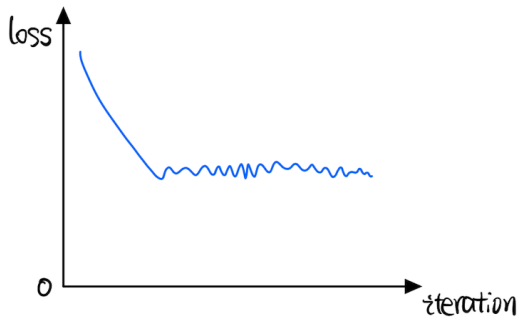


Figure 3: Training loss.

Circle the correct answer: Stochastic gradient descent OR Batch gradient descent

Explain briefly:

SOLUTION Stochastic Gradient Descent

Part (b) [2 MARKS]

Your friend is having had a hard time reducing the loss further. Could you come up with two solutions to help them further reduce the loss?

SOLUTION Reduce the learning rate in the end.

Use mini-batch gradient descent and increase the batch size.

Part (c) [1 MARK]

What is one advantage of using neural networks over linear regression?

SOLUTION A neural network with non-linear activation functions is a universal function approximator. It can model complex (especially non-linear) functions.

Part (d) [1 MARK]

What is one advantage of linear regression over neural networks?

SOLUTION Linear regression has a closed-form optimal solution. (b) Non-convex vs convex optimization - iterative optimization vs direct solution

Question 5. Naive Bayes [8 MARKS]

Let's continue with the candy example from the previous question. Suppose that the manufacture wants to give customers a little hint by wrapping the candies in **y**ellow or blue wrappers. The wrapper for each candy depends on the flavour probabilistically, but the conditional distribution is unknown to us.

Let $P_d \in [0, 1]$ denote the probability that a candy is dark chocolate. Let $P_{dy} \in [0, 1]$ denote the probability that a candy has a yellow wrapper given that the candy is dark chocolate. Let $P_{my} \in [0, 1]$ denote the probability that a candy has a yellow wrapper given that the candy is milk chocolate.

Part (a) [2 MARKS]

What is the likelihood of observing a milk chocolate candy in a blue wrapper?

SOLUTION

$$(1 - P_d)(1 - P_{my})$$

Part (b) [3 MARKS]

Your friend Avery unwrapped N candies. The flavour and wrapper counts are as follows.

- d of the N candies are dark chocolate.
  Among these d candies, d_y of the d wrappers are yellow and the rest are blue.
- $N - d$ of the N candies are milk chocolate.
  Among these $N - d$ candies, m_y of the $N - d$ wrappers are yellow and the rest are blue.

What is the likelihood of the data (the N candies)?

SOLUTION

$$\{P_d P_{dy}\}^{d_y}\{P_d(1 - P_{dy})\}^{d-d_y}\{(1 - P_d)P_{my}\}^{m_y}\{(1 - P_d)(1 - P_{my})\}^{N-d-m_y}$$

Part (c) [3 MARKS]

Derive the Maximum Likelihood estimates for P_d, P_{dy}, and P_{my}.

SOLUTION

$$P_d = d/N \tag{1}$$
$$P_{dy} = d_y/d \tag{2}$$
$$P_{my} = m_y/(N - d) \tag{3}$$