# Question Bank

**Disclaimer: Please note that this question bank is a useful resource for practice, there is no guarantee that questions on the exam will be derived solely from this material.**

### Questions on Linear Regression

Q1) Consider a data set in which each data point tn is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n \left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2$$

Find an expression for the solution w that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data-dependent noise variance and (ii) replicated data points.

Q2) Consider a linear basis function regression model for a multivariate target variable t having a Gaussian distribution of the form

$$p(\mathbf{t}|\mathbf{W}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{W}), \boldsymbol{\Sigma})$$

where

$$\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$$

together with a training data set comprising input basis vectors $\varphi(x_n)$ and corresponding target vectors tn, with n = 1,...,N. Show that the maximum likelihood solution $W_{ML}$ for the parameter matrix W has the property that each column is given by an expression of the form, which was the solution for an isotropic noise distribution. Note that this is independent of the covariance matrix Σ. Show that the maximum likelihood solution for Σ is given by

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)\left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)^{\mathrm{T}}$$

Q3) In the context of Bayesian linear regression, it is well understood that the uncertainty associated with the posterior distribution over model parameters decreases as the size of the dataset increases. This phenomenon can be mathematically analyzed using matrix identities.

Consider this matrix identity:

$$\left(\mathbf{M} + \mathbf{v}\mathbf{v}^{\mathrm{T}}\right)^{-1} = \mathbf{M}^{-1} - \frac{\left(\mathbf{M}^{-1}\mathbf{v}\right)\left(\mathbf{v}^{\mathrm{T}}\mathbf{M}^{-1}\right)}{1 + \mathbf{v}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{v}}$$

Here, M represents the prior covariance matrix, and v is a vector associated with the new data point that is being incorporated into the model. The goal of this question is to utilize this matrix identity to show how the uncertainty, denoted as $\sigma^2_N(x)$, behaves with increasing dataset size.

In the context of linear regression, the uncertainty $\sigma^2_N(x)$ is linked to the posterior distribution of the model parameters after observing N data points. Specifically, we are interested in showing that the uncertainty associated with the linear regression function given in the equation satisfies the following relationship:

$$\sigma^2_{N+1}(\mathbf{x}) \leqslant \sigma^2_N(\mathbf{x})$$

Q4) Consider a linear model of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

together with a sum-of-squares error function of the form

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

. Now suppose that Gaussian noise $\epsilon_i$ with zero mean and variance $\sigma^2$ is added independently to each of the input variables $x_i$. By making use of $E[i]=0$ and $E[\epsilon_i\epsilon_j] = \delta_{ij}\sigma^2$, show that minimizing $E_D$ averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter $w_0$ is omitted from the regularizer.

Q5) This problem involves simple linear regression without an intercept.

(a) Recall that the coefficient estimate $\hat{\beta}$ for the linear regression of Y onto X without an intercept. Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X?

(b) Generate an example in Python with n = 100 observations in which the coefficient estimate for the regression of X onto Y is different from the coefficient estimate for the regression of Y onto X.

(c) Generate an example in Python with n = 100 observations in which the coefficient estimate for the regression of X onto Y is the same as the coefficient estimate for the regression of Y onto X.

Q6) Suppose we run a ridge regression with parameter λ on a single variable X, and get coefficient a. We now include an exact copy $X^* = X$, and refit our ridge regression. Show that both coefficients are identical, and derive their value. Show in general that if m copies of a variable $X_j$ are included in a ridge regression, their coefficients are all the same

Q7) When the training set is small, the contribution of variance to error may be more than that of bias and in such a case, we may prefer a simple model even though we know that it is too simple for the task. Can you give an example?

Q8) Suppose you are using Polynomial Regression. You plot the learning curves and you notice that there is a large gap between the training error and the validation error. What is happening? What are three ways to solve this?

Q9) Suppose you use Batch Gradient Descent and you plot the validation error at every epoch. If you notice that the validation error consistently goes up, what is likely going on? How can you fix this?

Q10) Suppose you are using Ridge Regression, and you notice that the training error and the validation error are almost equal and fairly high. Would you say that the model suffers from high bias or high variance? Should you increase the regularization hyperparameter α or reduce it?

**References:**


- https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
- https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf
- https://dl.matlabyar.com/siavash/ML/Book/Ethem%20Alpaydin-Introduction%20to%20Machine%20Learning-The%20MIT%20Press%20(2014).pdf
- http://14.139.161.31/OddSem-0822-1122/Hands-On_Machine_Learning_with_Scikit-Learn-Keras-and-TensorFlow-2nd-Edition-Aurelien-Geron.pdf


---

## Questions on Perceptron

**1.** What are the values of weights `w0`, `w1`, and `w2` for the perceptron whose decision surface is illustrated in Figure 4.3? Assume the surface crosses the x1 axis at -1, and the x2 axis at 2.

**2.** Design a two-input perceptron that implements the boolean function `A ^ ~ B`. Design a two-layer network of perceptrons that implements `A XOR B`.

**3.** Consider two perceptrons defined by the threshold expression `w0 + w1*x1 + w2*x2 > 0`. Perceptron A has weight values

w0 = 1, w1 = 2, w2 = 1

, and perceptron B has weight values

w0 = 0, w1 = 2, w2 = 1

. True or false? Perceptron A is more general than perceptron B. (The concept of "more general than" is defined in Chapter 2).

**4.** Implement the delta training rule for a two-input linear unit. Train it to fit the target concept `-2 + x1 + 2*x2 > 0`. Plot the error `E` as a function of the number of training iterations. Plot the decision surface after 5, 10, 50, 100, ... iterations.

- (a) Try this using various constant values for `eta` (learning rate) and using a decaying learning rate of `eta0/i` for the i-th iteration. Which works better?
- (b) Try incremental and batch learning. Which converges more quickly? Consider both the number of weight updates and total execution time.

**5.** Derive a gradient descent training rule for a single unit with output `o`, where the output is given as a function of weighted inputs.

$$o = w_0 + w_1 x_1 + w_1 x_1^2 + \ldots + w_n x_n + w_n x_n^2$$

**6.** Explain informally why the delta training rule in Equation (4.10) is only an approximation to the true gradient descent rule of Equation (4.7).

**7.** Consider a two-layer feedforward artificial neural network (ANN) with two inputs a and b, one hidden unit `c`, and one output unit `d`. This network has five weights (`w_ac, w_bc, w_cd, w_dc, w_d0`), where `w_d0` represents the threshold weight for unit `d`. Initialize these weights to the values (0.1, 0.1, 0.1, 0.1, 0.1), then give their values after each of the first two training iterations of the backpropagation algorithm. Assume learning rate `eta = 0.3`, momentum `alpha = 0.9`, incremental weight updates, and the following training examples:

| a | b | d |
|---|---|---|
| 1 | 0 | 1 |
| 0 | 1 | 0 |

**8**. Revise the backpropagation algorithm in Table 4.2 so that it operates on units using the squashing function `tanh` in place of the sigmoid function. That is, assume the output of a single unit is `o = tanh(x)`. Give the weight update rule for output layer weights and hidden layer weights.

**9.** Recall the 8 x 3 x 8 network described in Figure 4.7. Consider trying to train an 8 x 1 x 8 network for the same task; that is, a network with just one hidden unit. Notice the eight training examples in Figure 4.7 could be represented by eight distinct values for the single hidden unit (e.g., 0.1, 0.2, ..., 0.8). Could a network with just one hidden unit therefore learn the identity function defined over these training examples?

**10.** Consider the alternative error function described in Section 4.8.1.

$$E(\vec{w}) \equiv \frac{1}{2} \sum_{d \in D} \sum_{k \in outputs} (t_{kd} - o_{kd})^2 + \gamma \sum_{i,j} w_{ji}^2$$

Derive the gradient descent update rule for this definition of E. Show that it can be implemented by multiplying each weight by some constant before performing the standard gradient descent update given in Table 4.2.

**Reference:**

Mitchell, T. M. (1997). *Machine Learning*. Available at: https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf

**Reference (FOR QUESTIONS 11 - 14):-**
https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf

**Q11** Give decision trees to represent the following boolean functions:

(a) A A -B

(b) A V [B A C]

(c) A XOR B

(d) [A A B] v [C A DI

**Q12** Consider the following set of training examples:

| Instance | Classification | $a_1$ | $a_2$ |
|---|---|---|---|
| 1 | + | T | T |
| 2 | + | T | T |
| 3 | − | T | F |
| 4 | + | F | F |
| 5 | − | F | T |
| 6 | − | F | T |

(a) What is the entropy of this collection of training examples with respect to the target function classification?

(b) What is the information gain of a2 relative to these training examples?

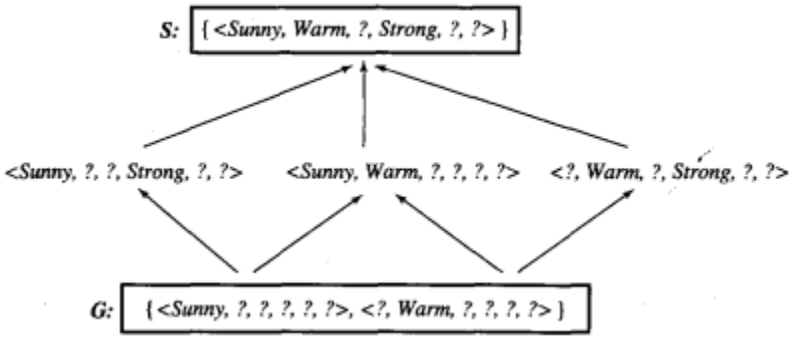
**Q13** True or false: If decision tree D2 is an elaboration of tree DI, then DI is more general-than D2. Assume DI and D2 are decision trees representing arbitrary boolean functions, and that D2 is an elaboration of DI if ID3 could extend DI into D2. If true, give a proof; if false, a counterexample

**Q14.** ID3 searches for just one consistent hypothesis, whereas the CANDIDATEELIMINATION algorithm finds all consistent hypotheses. Consider the correspondence between these two learning algorithms.

(a) Show the decision tree that would be learned by ID3 assuming it is given the four training examples for the Enjoy Sport?

| Example | *Sky* | *AirTemp* | *Humidity* | *Wind* | *Water* | *Forecast* | *EnjoySport* |
|---|---|---|---|---|---|---|---|
| 1 | Sunny | Warm | Normal | Strong | Warm | Same | Yes |
| 2 | Sunny | Warm | High | Strong | Warm | Same | Yes |
| 3 | Rainy | Cold | High | Strong | Warm | Change | No |
| 4 | Sunny | Warm | High | Strong | Cool | Change | Yes |

(b) What is the relationship between the learned decision tree and the version space




that is learned from these same examples? Is the learned tree equivalent to one of the members of the version space?

(c) Add the following training example, and compute the new decision tree. This time, show the value of the information gain for each candidate attribute at each step in growing the tree.

Sky Air-Temp Humidity Wind Water Forecast Enjoy-Sport?

Sunny Warm Normal Weak Warm Same No

(d) Suppose we wish to design a learner that (like ID3) searches a space of decision tree hypotheses and (like CANDIDATE-ELIMINATION) finds all hypotheses consistent with the data. In short, we wish to apply the CANDIDATE-ELIMINATION algorithm to searching the space of decision tree hypotheses. Show the S and G sets that result from the first training example from Table in a part. Note S must contain the most specific decision trees consistent with the data, whereas G must contain the most general. Show how the S and G sets are refined by the second training example (you may omit syntactically distinct trees that describe the same concept). What difficulties do you foresee in applying CANDIDATE-ELIMINATION to a decision tree hypothesis space?

**Reference (FOR QUESTION 15):-**
https://courses.cs.washington.edu/courses/cse546/14au/exams/14au_midterm_sol.pdf

**Q15**

We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied.

| GPA | Studied | Passed |
|---|---|---|
| L | F | F |
| L | T | T |
| M | F | F |
| M | T | T |
| H | F | T |
| H | T | T |

For this problem, you can write your answers using log2 , but it may be helpful to note that log2 3 ≈ 1.6.

1. [4 points] What is the entropy H(Passed)?

2. [4 points] What is the entropy H(Passed | GPA)?

3. [4 points] What is the entropy H(Passed | Studied)?

4. [4 points] Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations.

**Reference (FOR QUESTIONS 16 - 21 )**:-
https://www.cs.cmu.edu/~tom/10701_sp11/midterm.pdf

**Q16.** [2 pts] If we train a Naive Bayes classifier using infinite training data that satisfies all of its modeling assumptions (e.g., conditional independence), then it will achieve zero training error over these training examples.

**Q17.** [2 pts] If we train a Naive Bayes classifier using infinite training data that satisfies all of its modeling assumptions (e.g., conditional independence), then it will achieve zero true error over test examples drawn from this same distribution.

**Q18**

Consider a binary classification problem with variable X1 ∈ {0, 1} and label Y ∈ {0, 1}. The true generative distribution P(X1, Y ) = P(Y )P(X1|Y ) is shown as Table 1 and Table 2. Question [4 pts]: Now suppose we have trained a Naive Bayes classifier, using infinite training data generated according to Table 1 and Table 2. In Table 3, please write down the predictions from the trained Naive Bayes for different configurations of X1. Note that Yˆ (X1) in the table is the

decision about the value of Y given X1. For decision terms in the table, write down either Yˆ = 0 or Yˆ = 1; for probability terms in the table, write down the actual values (and the calculation process if you prefer, e.g., 0.8 ∗ 0.7 = 0.56).

Table 1: $P(Y)$

| $Y=0$ | $Y=1$ |
|---|---|
| 0.8 | 0.2 |

Table 2: $P(X_1|Y)$

| | $X_1=0$ | $X_1=1$ |
|---|---|---|
| $Y=0$ | 0.7 | 0.3 |
| $Y=1$ | 0.3 | 0.7 |

Table 3: Predictions from the trained Naive Bayes

| | $\hat{P}(X_1, Y=0)$ | $\hat{P}(X_1, Y=1)$ | $\hat{Y}(X_1)$ |
|---|---|---|---|
| $X_1=0$ | | | |
| $X_1=1$ | | | |

**Question** [3 pts]: What is the expected error rate of this Naive Bayes classifier on testing examples that are generated according to Table 1 and Table 2? In other words, P(Yˆ (X1) 6= Y ) when (X1, Y ) is generated according to the two tables. Hint: P(Yˆ (X1) 6= Y ) = P(Yˆ (X1) 6= Y, X1 = 0) + P(Yˆ (X1) 6= Y, X1 = 1).

## QUESTIONS Description

Consider a classification problem with two boolean variables X1, X2 ∈ {0, 1} and label Y ∈ {0, 1}. In Figure 1 we show two positive (“+”) and two negative (“-”) examples.

**Q19**. [2 pts]: Draw (or just simply describe) a decision tree that can perfectly classify the four examples in Figure 1.

**Q20**. [3 pts]: In the class we learned the training rule to grow a decision tree: we start from a single root node and iteratively split each node using the “best” attribute selected by maximizing the information gain of the split. We will stop splitting a node if: 1) examples in the node are already pure; or 2) we cannot find any single attribute that gives a split with positive information gain. If we apply this training rule to the examples in Figure 1, will we get a decision tree that perfectly classifies the examples? Briefly explain what will happen.

**Q21.** [5 pts]: Suppose we learn a Naive Bayes classifier from the examples in Figure 1, using MLE (maximum likelihood estimation) as the training rule. Write down all the parameters and their estimated values (note: both P(Y ) and P(Xi |Y ) should be Bernoulli distributions). Also, does this learned Naive Bayes perfectly classify the four examples?

**Q22.** [3 pts]: Show that the 'tanh' function and the logistic sigmoid function are related by tanh(a)=2σ(2a) − 1.*Reference*: "Pattern Recognition and Machine Learning" by Christopher M. Bishop, Chapter 4.1 - Perceptron".

**Q23.** [3 pts] Illustrate the geometrical interpretation of the decision boundary produced by the Perceptron. How does this relate to the dot product? *Reference*: "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Chapter 6 - Deep Feedforward Networks.

**Q24**. [5 pts] Compare and contrast the Perceptron with logistic regression. What are the key differences in terms of loss function and decision boundary? *Reference*: "Pattern Recognition and Machine Learning" by Christopher M. Bishop, Chapter 4.3 - Perceptron vs. Logistic Regression.

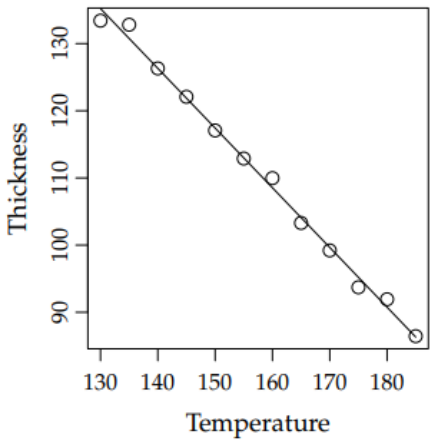
**Q25.** [5 pts+] Show that the Perceptron algorithm updates the weight vector by projecting the input vector onto the hyperplane defined by the current decision boundary. Derive the update rule mathematically**.** *Reference*: "Pattern Recognition and Machine Learning" by Christopher M. Bishop, Chapter 4.1 - Perceptron.

**Q26.** [5 pts+] Prove the Perceptron convergence theorem: If the data is linearly separable, show that the Perceptron algorithm converges in a finite number of steps. Provide a mathematical proof and specify the conditions for convergence. *Reference*: "Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, Chapter 12 - Linear Methods for Classification.

## Questions on Linear models for regression

**Reference 22:** [https://02323.compute.dtu.dk/enotes/solutions-chapter5.pdf](https://02323.compute.dtu.dk/enotes/solutions-chapter5.pdf)

**Question 22.** On a machine that folds plastic film the temperature may be varied in the range of 130-185°C. For obtaining, if possible, a model for the influence of temperature on the folding thickness, n = 12 related set of values of temperature and the fold thickness were measured that is illustrated in the following figure:




a) Determine by looking at the figure, which of the following sets of estimates for the parameters in the usual regression model is correct:

1) $\hat{\beta}_0 = 0, \hat{\beta}_1 = -0.9, \hat{\sigma} = 36$
2) $\hat{\beta}_0 = 0, \hat{\beta}_1 = 0.9, \hat{\sigma} = 3.6$
3) $\hat{\beta}_0 = 252, \hat{\beta}_1 = -0.9, \hat{\sigma} = 3.6$
4) $\hat{\beta}_0 = -252, \hat{\beta}_1 = -0.9, \hat{\sigma} = 36$
5) $\hat{\beta}_0 = 252, \hat{\beta}_1 = -0.9, \hat{\sigma} = 36$

b) What is the only possible correct answer:
1) The proportion of explained variation is 50% and the correlation is 0.98
2) The proportion of explained variation is 0% and the correlation is −0.98
3) The proportion of explained variation is 96% and the correlation is −1
4) The proportion of explained variation is 96% and the correlation is 0.98
5) The proportion of explained variation is 96% and the correlation is −0.98

**Reference 23-26:**[https://www.toothillschool.co.uk/data/files/dept/maths/s1_regression.pdf](https://www.toothillschool.co.uk/data/files/dept/maths/s1_regression.pdf)

**Question 23.** A biologist assumes that there is a linear relationship between the amount of fertilizer supplied to tomato plants and the subsequent yield of tomatoes obtained.

Eight tomato plants of the same variety were selected at random and treated, weekly, with a solution in which x grams of fertilizer was dissolved in a fixed quantity of water. The yield, y kilograms, of tomatoes was recorded.

| Plant | *A* | *B* | *C* | *D* | *E* | *F* | *G* | *H* |
|---|---|---|---|---|---|---|---|---|
| *x* | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 |
| *y* | 3.9 | 4.4 | 5.8 | 6.6 | 7.0 | 7.1 | 7.3 | 7.7 |

(a) Plot a scatter diagram of yield, y, against amount of fertilizer, x.

(b) Calculate the equation of the least squares regression line of y on x.

**Question 24.** Over a period of one year, a greengrocer sells tomatoes at six different prices (x pence per kilogram). He calculates the average number of kilograms, y, sold per day at each of the six different prices. From these data the following are calculated.

$\Sigma x = 200$ $\Sigma y = 436$ $\Sigma xy = 12\,515$

$\Sigma x^2 = 7250$ $\Sigma y^2 = 39\,234$ $n = 6$

(a) Calculate the value of the product moment correlation coefficient.

**Question 25.** The table shows a Verbal Reasoning test score, x, and an English test score, y, for each of a random sample of 8 children who took both tests.

| **Child** | *A* | *B* | *C* | *D* | *E* | *F* | *G* | *H* |
|---|---|---|---|---|---|---|---|---|
| ***x*** | 112 | 113 | 110 | 113 | 112 | 114 | 109 | 113 |
| ***y*** | 69 | 65 | 75 | 70 | 70 | 75 | 68 | 76 |

(a) Calculate the value of the product moment correlation coefficient between the scores in Verbal Reasoning and English.

(b) Comment briefly, in context, on the result obtained in part (a).

**Question 26.** Nasser organizes a street collection for a mental health charity. The collection takes place in a large city on a particular Saturday. Volunteers, with collecting tins, stand in busy places and ask passers-by for donations. The following table shows, for ten volunteers, the times, x minutes, they spent collecting together with the amounts, to the nearest pound, y, they collected.

| **Collector** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
|---|---|---|---|---|---|---|---|---|---|---|
| ***x*** | 65 | 187 | 126 | 52 | 143 | 90 | 157 | 74 | 88 | 195 |
| ***y*** | 21 | 55 | 23 | 8 | 28 | 27 | 44 | 19 | 17 | 47 |

(a) Plot a scatter diagram of the data.

(b) Calculate the equation of the regression line of y on x and draw the line on your scatter diagram.

(c) The following table shows the residuals for some of the collectors.

| Collector | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Residual | 6.25 | 7.50 | −8.13 | −3.26 | −7.69 | 5.54 | 4.55 | 1.83 | | |

(i) Calculate the residuals for collectors I and J.

(ii) Calculate the mean magnitude of the ten residuals.

**Reference 27:** https://www.gdcollegebegusarai.com/course_materials/july/eco184.pdf

**Question 27.** The sales of a company (in million dollars) for each year are shown in the table below.

| x (year) | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|
| y (sales) | 12 | 19 | 29 | 37 | 45 |

a) Find the least square regression line y = a x + b.

b) Use the least squares regression line as a model to estimate the sales of the company in 2012.

**Reference 28-29:** https://www.gdcollegebegusarai.com/course_materials/july/eco184.pdf

**Question 28.** A student who waits on tables at a restaurant recorded the cost of meals and the tip left by single diners.

| Meal Cost | $4.75 | $6.84 | $12.52 | $20.42 | $8.97 |
|---|---|---|---|---|---|
| Tip | $0.50 | $0.90 | $1.50 | $3.00 | $1.00 |

If the next diner orders a meal costing $10.50, how much tip should the waiter expect to receive?

Equation_____ Tip expected_____

**Question 29.** The table below gives the number of hours spent studying for a science exam (x) and the final exam grade (y).

| X | 2 | 5 | 1 | 0 | 4 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| Y | 77 | 92 | 70 | 63 | 90 | 75 | 84 |

Predict the exam grade of a student who studied for 6 hours.

Equation_____ Grade expected_____

**Questions on Decision Trees**

**Reference 30-31:**
https://www.researchgate.net/publication/370591757_Question_bank_on_decision_tree_algorithm

**Question 30.** A decision tree classifier is used to predict the failure mode of a boiler system with a dataset of 300 samples, where 200 samples indicate normal operation and 100 samples indicate tube leakage. What is the entropy of the classifier?

**Question 31.** A car manufacturer is testing a new engine design that can produce three types of failures: overheating, oil leakage, and poor fuel efficiency. Data has been collected on 500 engines produced by the new design, and has labeled each engine as either failure-free, overheating, oil leakage, or poor fuel efficiency. The data is summarized in the following table:

| Engine status | Number of engines |
|---|---|
| Failure-free | 350 |
| Overheating | 100 |
| Oil leakage | 30 |
| Poor fuel efficiency | 20 |

What is the entropy of the dataset?

---

Q32) Question: Explain the principle of the gradient descent algorithm. Accompany your explanation with a diagram. Explain the use of all the terms and constants that you introduce and comment on the range of values that they can take.(Gradient descent)

Q33) Question: Consider the following set of training examples:

| Instance | Classification | a1 | a2 |
|---|---|---|---|
| 1 | + | T | T |
| 2 | + | T | T |
| 3 | - | T | F |
| 4 | + | F | F |
| 5 | - | F | T |
| 6 | - | F | T |

What is the information gain of a2 relative to these training examples? Provide the equation for calculating the information gain as well as the intermediate results.

---

Q1. Indicate (by circling) how the bias and variance change in response to the action: Reduce the number of leaves in a decision tree:

| Bias | Variance |
|---|---|
| Increase | Increase |
| Decrease | Decrease |
| No Change | No Change |

Q2.Consider a Multi-Layer Perceptron (MLP) model with one hidden layer and one output layer. The hidden layer has 10 neurons, and the output layer has 3 neurons. The input to the MLP is a 5-dimensional vector. Each neuron is connected to every neuron in the previous layer, and a bias term is included for each neuron. The activation function used is the sigmoid function. Calculate the total number of trainable parameters in this MLP model.

Q3.[True/False] The depth of a learned decision tree can be larger than the number of training examples used to create the tree.

Q4.Consider k-fold cross-validation. Let's consider the tradeoffs of larger or smaller k (the number of folds).
Please select one of the multiple-choice options.
With a higher number of folds, the estimated error will be, on average,
(a) Higher.
(b) Lower.
(c) Same.
(d) Can't tell

Q5. Suppose you have picked the parameter θ for a model using 10-fold cross-validation. The best way to pick a final model to use and estimate its error is to
(a) pick any of the 10 models you built for your model; use its error estimate on the held-out data
(b) pick any of the 10 models you built for your model; use the average CV error for the 10 models as its error estimate

(c) average all of the 10 models you got; use the average CV error as its error estimate
(d) average all of the 10 models you got; use the error the combined model gives on the full training set
(e) train a new model on the full data set, using the θ you found; use the average CV error as its error estimate

Q6.Suppose we want to compute 10-Fold Cross-Validation error on 100 training examples. We need to compute error N1 times, and the Cross-Validation error is the average of the errors. To compute each error, we need to build a model with data of size N2, and test the model on the data of size N3. What are the appropriate numbers for N1, N2, N3?
(a) N1 = 10, N2 = 90, N3 = 10
(b) N1 = 1, N2 = 90, N3 = 10
(c) N1 = 10, N2 = 100, N3 = 10
(d) N1 = 10, N2 = 100, N3 = 100

Q7.Derive the variance formula

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

This appears to fail if ρ is negative; diagnose the problem in this case.

References:

https://www.cs.cmu.edu/~ggordon/10601/exams/midterm/midterm_sol.pdf

https://gateoverflow.in/413584/gate-science-artificial-intelligence-sample-paper-question
https://www.cs.cmu.edu/~aarti/Class/10701/exams/midterm2010f_sol.pdf
https://www.cs.cmu.edu/~epxing/Class/10701/exams/12f-601-midterm.pdf
5,6 https://www.seas.upenn.edu/~cis520/exams/midterm_2016_solns.pdf

---

Q1) Explain the concept of the credit-assignment problem in the context of MLPs and how backpropagation helps solve it.

Q2) What is the benefit of using antisymmetric activation functions (e.g., hyperbolic tangent) compared to non-symmetric functions in MLPs?

Q3) Briefly describe the importance of normalizing input data before feeding it into a neural network.

Q4) Why is it important to use examples with the largest training error when maximizing the information content during backpropagation training?

Q5) How are the initial weights generally chosen in an MLP?

Q6) What is a major downside of assigning large initial values to synaptic weights?

Q7) Which of the following describes a practical use case of the Generalized Delta Rule?
a) Assigns constant learning rate throughout the training
b) Uses momentum constant to smooth out weight updates
c) Ignores the weight correction during backpropagation
d) Stops learning once the local gradient is zero

Answer: B)

Q8) Which of the following is NOT a phase in the functioning of an MLP?
a) Forward phase
b) Backward phase
c) Lateral phase
d) Credit-assignment phase

Ans: C)

Q9) Imagine you are training a neural network, but the training loss plateaus after several iterations. Describe a systematic approach to diagnose and resolve this issue. Include discussions on learning rate, initialization, architecture, and other hyperparameters that could be adjusted to improve learning.

Q10) In the context of backpropagation, derive the mathematical expression for the local gradient at a hidden neuron. Show how the chain rule of calculus is applied and explain how this expression is used to update the weights during training.

---

Questions:

Q1- Consider the problem of predicting whether the university will be closed on a particular day. We will assume that the factors which decide this are whether there is a snowstorm, a weekend, or an official holiday. Suppose we have the training examples described in the table.

| Snowstorm | Holiday | Weekend | Closed |
|---|---|---|---|
| T | T | F | F |

| | | | |
|---|---|---|---|
| T | T | F | T |
| F | T | F | F |
| T | T | F | F |
| F | F | F | F |
| F | F | F | T |
| T | F | F | T |
| F | F | F | T |

1. What would be the effect of the Weekend attribute on the decision tree if it were made the root? Explain in terms of information gain.
2. If we cannot make Weekend the root node, which attribute should be the root node of the decision tree? Explain your reasoning and show your calculations. (You may use $\log_2 0.75 = -0.4$ and $\log_2 0.25 = -2$)

Q2 - Draw a simple MLP network by using linear activation function to get the desired output y from the inputs x1 and x2. Check truth table 1 to understand the mathematical realtion between them. Assign proper weights and threshold values to the the MLP network. You can use maximum of 3 layers (excluding the input layer). Note: x1, x2 and y are binary variables.

| $x_1$ | $x_2$ | $\mathbf{y}$ |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

Q3 - Compare between Random Forest and Bootstrap Aggregation, which works better for bias-variance trade-off and why?

Q4 - A logistic regression model is facing the problem of high bias. The problem can be resolved by introducing l2 regularization. True or False, and why?

Q5 - I have a classification problem in R2, with input features X1, X2, and target variable Y . Also, X1, X2, Y ∈ {0, 1}. The Y is defined as:

Y = 1 if X1!= X2

Y= 0 if X1== X2

I am using logistic regression to classify the above dataset perfectly in R2. Is it possible, and why?

Q6- Find out the precision, recall and F1 score for the following true and predicted labels. (Both are arranged in order)
Note: For F1 score, use balanced F
true = {True, False, True, False, True, False, True, False}
predicted = {True, True, True, True, True, True, True, False}

Q7- What is the loss function used to train a logistic regression model? Define the probabilistic model used, and write the mathematical expression for the loss (log-likelihood).

Q8- There is a very rare disease where a person feels the urge to eat non-edible objects. Dr. Hussain has collected a dataset having various attributes for detecting this disease in humans. He then built a classifier using ML techniques which achieved a training accuracy of 88% and validation accuracy of 94%. Is accuracy the right metric for this problem? If not, which metric should he use?
Q9 - Kaira's ML model is giving 99% accuracy on train set but only 34% accuracy on the test set. Can collecting more training data help with this issue?

Q10 - For a multi-class classification problem, what are the maximum and minimum values possible for Shanon's entropy? When are these values achieved?

Q11 - Suppose Arya has trained a linear regression model and plotted the residual (y actual − y predicted) and predicted values on a plane. She observes that there is a relationship between them. What can she say about the model trained ?
1. The model is overfitted.
2. The model is underfitted.
3. The model has failed to capture the relation between input data and output completely.
4. It is not possible to comment as the information provided in the question is incomplete.

Q12 - When and why stochastic gradient descent is preferred over batch gradient descent?

Q13 - Consider we are performing Random Forests based on classification data, and the relative frequencies of the class you are observing in the dataset are 0.65, 0.35, 0.29, and 0.5. What is the Gini index?

a) 0.423
b) 0.084
c) 0.12
d) 0.25

Q14 - **Statement:** For a one vs rest multi-classification scheme, we need n(n−1)/2 classifiers to classify n classes. Specify True or False with proper reason.

Q15- **Statement:** A logistic regression model uses the logit function to handle nonlinear boundaries in the data. Specify True or False with proper reason.

Q16 - What is the main purpose of an activation function in a Multi-Layer Perceptron (MLP)?

a) To initialize the weights

b) To introduce non-linearity into the model

c) To calculate the output layer's value

d) To improve the gradient during backpropagation

Q17 - In an MLP, the weight update rule during backpropagation is influenced by:

a) The learning rate and the gradient of the loss function
b) The number of neurons in the input layer
c) The type of activation function used in the output layer
d) The size of the dataset

Q18- If the log-odds (logit) in a logistic regression model is given as logit(p)=1.5, what is the predicted probability p?

a) 0.43
b) 0.58
c) 0.82
d) 0.85

Q19- In simple linear regression, the relationship between the dependent variable y and independent variable x is given by:

$y=\beta 0+\beta 1x+\epsilon$

Given the data points (1,2), (2,3), and (3,5) compute the slope β1 of the best-fit line.

a) 0.5
b) 1.0
c) 1.5
d) 2.0

References:1. https://www.cs.cmu.edu/~mgormley/courses/10601-f19/slides/lecture10-mid.pdf
2. https://www.sanfoundry.com/machine-learning-questions-answers-random-forest-algorithm/

Q20-You are training a logistic regression model. You initialize the parameters with 0's. Is this a good idea? Explain your answer.

---

Q1. Explain the role of regularization techniques like dropout, weight decay, and batch normalization in neural networks. For each technique, provide an example scenario where it would be beneficial and explain the underlying mechanism that helps prevent overfitting.

Q2. Compare and contrast different activation functions (Sigmoid, Tanh, ReLU, and Leaky ReLU) in terms of their advantages, disadvantages, and areas of application. Provide mathematical expressions for each function and their derivatives, and discuss in which scenarios one function might outperform the others.

Q3. Consider the XOR problem, which is linearly inseparable. Design an MLP with appropriate layers and activation functions to solve the XOR problem. Explain why an MLP with at least one hidden layer is required and show step-by-step how backpropagation would update the weights for a given input.

Q4. Consider a scenario where you are tasked with designing a deep neural network for a time-series prediction problem. Discuss how you would choose the network architecture (number of layers, neurons per layer, and activation functions). Additionally, describe how you would tune hyperparameters such as learning rate and batch size, and what considerations you would take for training stability and convergence.

Q5. In the context of gradient descent, explain how the vanishing gradient problem occurs in MLPs using the sigmoid activation function.

Q6. What is the derivative of the hyperbolic tangent (tanh) activation function, and why is it preferred over the sigmoid in some cases?

Q7. What is a major downside of assigning large initial values to synaptic weights?
a) Slow learning rate
b) Neurons may go into saturation
c) Neurons may fail to converge
d) The network may oscillate

Q8) You are training a very deep feedforward neural network with 100 layers using the sigmoid activation function. After several iterations of training, you notice that the gradients in the lower layers (closer to the input) have become excessively large, leading to gradient explosion, while the upper layers have very small gradients.

1. Explain why the gradients are exploding in the lower layers of the network. Use mathematical reasoning involving the sigmoid function.
2. Derive a potential fix using a weight initialization scheme and explain how this will help reduce the problem. Use detailed steps in your derivation.
3. Propose an alternative activation function and explain why it might help prevent gradient explosion in such a deep network.

Q9) You are building a neural network to predict whether a patient has a rare disease. In this case, a false negative (i.e., predicting that a patient does not have the disease when they actually do) is far more costly than a false positive. You want to design a custom cost function to reflect this asymmetry.

1. Formulate a custom loss function that penalizes false negatives more heavily than false positives. Express your loss function mathematically
2. Explain how this loss function will affect the weight updates during training. Provide a detailed explanation of how the gradient will change compared to a standard loss function like binary cross-entropy.
3. Propose a training strategy or adjustment (such as in the learning rate) that would complement this loss function for better convergence.

Q10) You are tasked with designing a neural network to classify images into one of five categories: A, B, C, D, and E. However, the training data is highly imbalanced, with the majority of the images belonging to class A and very few images belonging to class E. Standard neural networks trained on this data tend to perform poorly, especially on the minority class (class E).

1. Propose a modified loss function or strategy to handle this class imbalance. Explain how your approach works mathematically and why it will help the network perform better on minority classes.
2. Suppose you implement a softmax output layer and cross-entropy loss function for this multi-class classification problem. How would you adjust the softmax probabilities or the loss calculation to ensure the network learns well for minority classes? Show the relevant equations.
3. Discuss a training approach (other than modifying the loss function) to improve the network's performance on minority classes, such as data augmentation or resampling techniques. Explain how these approaches help and the potential trade-offs involved.

**Solutions to Q8,9,10:**
https://drive.google.com/file/d/12aCh8H6Hsx82YxaW_vD6Y4WQZ7AVoJXe/view?usp=sharing

Q11)Suppose you design a multilayer perceptron for classification with the following architecture. It has a single hidden layer with the hard threshold activation function. The output layer uses the softmax activation function with cross-entropy loss. What will go wrong if you try to train this network using gradient descent? Justify your answer in terms of the backpropagation rules.

---

Q1. Consider a model with high variance and low bias.

(a) Explain what high variance and low bias mean in the context of model performance.

(b) Discuss two strategies to address high variance while maintaining model performance.

Q2. In logistic regression, the model predicts the probability of the positive class using the sigmoid function $\sigma(z)$, where $z = \beta_0 + \beta_1 x$.

(a) Derive the log-odds (logit) function from the sigmoid function
(b) Discuss how the choice of threshold affects the precision and recall of the logistic regression model

Q3. Given $\sigma(x) = 1/3$, where $\sigma(a)$ is the sigmoid function, compute the gradient of the negative of the sigmoid function, i.e. $\sigma'(-x)$.

Q4. Suppose you build a linear regression model, and upon examining the residuals, you notice a clear pattern rather than random noise. What might this indicate about your model?

Q5. Consider an MLP where all weights are initialized to the same value. Provide a theoretical explanation for why this weight initialization scheme will fail to train the network properly. Prove that the gradient descent update for all weights will be identical at each step, and thus the network will fail to break symmetry.

Q6.You are training a logistic regression model using gradient descent on a binary classification problem.
(a) Derive the gradient descent update rule for the weight vector $\theta$ with respect to the negative log-likelihood loss function.
(b) Suppose L2 regularization is applied to prevent overfitting. Modify the update rule to incorporate L2 regularization. What effect does this have on the weights over time?
Q7. A machine learning practitioner is trying to balance bias and variance while training an MLP model with several hidden layers on a small dataset.
(a) Define the bias-variance trade-off in the context of MLP training.
(b) How does increasing the depth of the network (number of hidden layers) affect bias and variance? Suggest strategies to mitigate overfitting in this context.

Q8. Consider an MLP with a large number of hidden layers trained using gradient descent.
(a) Derive the gradient descent update rule for the weights of a hidden layer in an MLP.
(b) Discuss why gradient descent might converge more slowly in deep MLPs compared to shallow networks or logistic regression. How can techniques like momentum or adaptive learning rates improve convergence?

Q9.

Consider an MLP with a single hidden layer, input $X \in \mathbb{R}^{n \times p}$, hidden layer weights $W_1 \in \mathbb{R}^{p \times h}$, and output weights $W_2 \in \mathbb{R}^{h \times 1}$. The activation function used is ReLU, and the output layer has a sigmoid activation function.

(a) Write down the expression for the output $\hat{y}$ as a function of $X$, $W_1$, and $W_2$.

(b) Derive the gradient of the output with respect to $W_1$ and $W_2$ for the binary cross-entropy loss function.

(c) Now, introduce L2 regularization on the weights. Derive the update rule for both $W_1$ and $W_2$ using gradient descent with L2 regularization.

Q10.

In linear regression, let $\hat{y}(x)$ be the prediction of the model for input $x$. The expected prediction error can be decomposed as follows:

$$\mathbb{E}[(y - \hat{y}(x))^2] = \text{Bias}(\hat{y}(x))^2 + \text{Var}(\hat{y}(x)) + \sigma^2$$

where $\sigma^2$ is the irreducible error.

(a) Derive the bias-variance decomposition for linear regression, starting from the definition of the mean squared error.

(b) Explain how regularization (e.g., Ridge regression) affects the bias and variance components of the error.

---

Q1.

Explain the gradient descent method in the context of linear regression and how the learning rate affects convergence. Derive the update rule for the weight vector θ using gradient descent to minimise the Mean Squared Error (MSE) loss function.

Now, suppose the gradient descent algorithm is modified to include an L2 regularization term (Ridge Regression), which penalizes large weights to prevent overfitting. The regularized cost function is given by $J_{\text{reg}}(\theta) = J(\theta) + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2$, where θ is the regularization parameter. Derive the new update rule for the weight vector θ incorporating the L2 regularization term.

Q2

You are tasked with developing a machine learning model to predict the likelihood of machine failures in an industrial plant within the next 30 days. The dataset used for this model includes

various features like temperature, vibration levels, and historical maintenance records. After training the model, it was tested on a validation set consisting of 1,000 machines. The results are as follows:

- 240 machines actually failed within 30 days. Your model correctly predicted 180 of these failures.
- 60 machines that eventually failed were not flagged by the model as likely to fail.
- Out of the 760 machines that did not fail, the model incorrectly predicted that 120 would fail.

Additional Information:

1. Cost of Maintenance: Each machine flagged as likely to fail underwent immediate maintenance, costing the plant $5,000 per machine.
2. Cost of Machine Failure: Each actual machine failure costs the company $50,000 in damages and lost productivity.

Based on this information:

1. Construct the confusion matrix for this classification problem.
2. Calculate the precision and recall for the “Failure” class.
3. Calculate the money lost on unnecessary maintenance and machine failuress. Based on this analysis, determine whether the factory should prioritize improving precision or recall.

Q3.

Consider a dataset with input features X $\in \mathbb{R}^{n \times p}$ and target vector y $\in \mathbb{R}^{n}$ , where n > p . The goal is to fit a linear regression model $y = X\theta + \epsilon$ , where $\theta \in \mathbb{R}^{p}$ is the parameter vector and \epsilon is the error term.

a) Derive the normal equations for the least squares estimate of \theta and show that the solution is given by $\theta = (X^{\top}X)^{-1}X^{\top}y$ .

b) Explain the geometric interpretation of the least squares solution in terms of projecting y onto the column space of X .

c) Discuss how multicollinearity among the columns of X can affect the stability of the solution and suggest methods to address this issue.

Q4

A dataset contains observations $\{(x_i, y_i)\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^{p}$ and $y_i \in \{0, 1\}$ . The probability that $y_i = 1$ is modeled using logistic regression:

$$P(y_i = 1|x_i) = \sigma(\theta^\top x_i) = \frac{1}{1 + e^{-\theta^\top x_i}}$$

a) Derive the log-likelihood function $\ell(\theta)$ for the logistic regression model.

b) Compute the gradient $\nabla_\theta \ell(\theta)$ of the log-likelihood function with respect to $\theta$ .

c) Explain how gradient ascent can be used to find the maximum likelihood estimate of $\theta$ and discuss the convergence properties of this method.

Q5

A machine learning practitioner is fitting polynomial regression models of varying degrees d to a dataset with input x and output y .

a) Define bias and variance in the context of model prediction error and explain how they relate to model complexity.

b) Describe how k -fold cross-validation can be used to select the optimal polynomial degree d to balance bias and variance.

Q6

Given a small dataset with n = 50 observations, a researcher wants to estimate the accuracy of a predictive model.

a) Explain the bootstrap resampling technique and how it can be used to estimate the sampling distribution of a statistic.

b) Describe how to compute bootstrap confidence intervals for the model's prediction error.

c) Discuss the limitations of bootstrap methods in the context of small sample sizes and dependent data

Q7

Consider linear regression with a large number of features p , where some features may be irrelevant.

a) Compare and contrast Ridge Regression (L2 regularization) and Lasso Regression (L1 regularization) in terms of their effects on coefficient estimates.

b) Explain how Lasso Regression can perform variable selection and discuss scenarios where this property is particularly useful.

Q8

A dataset has $n = 100$ observations and $p = 500$ features. A linear regression model is fitted to predict the target variable.

a) Explain why models in high-dimensional settings are prone to overfitting and how this relates to the bias-variance trade-off.

b) Propose methods to reduce overfitting in this context, such as dimensionality reduction or regularization.

Q9

A logistic regression model is developed to detect fraudulent transactions. The confusion matrix on a test set is as follows:

- True Positives (TP): 50
- False Positives (FP): 150
- True Negatives (TN): 9,800
- False Negatives (FN): 0

a) Calculate the precision, recall, and F1-score of the model.

b) Explain why accuracy may not be an appropriate metric in this scenario and discuss the importance of precision and recall.

c) Describe how adjusting the classification threshold can improve the model's performance and discuss the trade-offs involved.

Q10

In a linear regression problem where the number of features exceeds the number of observations ( $p > n$ ), the ordinary least squares solution is not unique.

a) Explain why the least squares solution is not unique in this case.

b) Describe how Ridge Regression (L2 regularization) can provide a unique solution and discuss the effect of the regularization parameter $\lambda$ .

**Q11.** [5 pts] Derive the bias-variance decomposition for the expected mean squared error (MSE) of a model f(x)f(x)f(x). Clearly define bias and variance, and show that:

$$\mathbb{E}_{x,y}[(y - f(x))^2] = (\text{Bias}^2) + (\text{Variance}) + \sigma^2$$

where σ2\sigma^2σ2 is the irreducible error (noise). Prove each step of the decomposition. *Reference*: "The Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, Chapter 7 - Model Assessment and Selection

**Q12.** Mathematically analyze the relationship between bias and variance for polynomial regression. For a dataset sampled from a quadratic function with Gaussian noise, derive the bias and variance of the learned polynomial model as a function of the polynomial degree. Explain how model complexity influences both bias and variance. *Reference*: "Machine Learning: A Probabilistic Perspective" by Kevin P. Murphy, Chapter 7 - Bias-Variance Tradeoff.

**Q13**.Why do simpler models generally exhibit higher bias and lower variance compared to more complex models? Discuss with examples. *Reference*: "Machine Learning: A Probabilistic Perspective" by Kevin P. Murphy, Chapter 7 - Bias-Variance Tradeoff.

**Q14**. Provide an example of a high-bias model and a high-variance model. How do they perform differently on the training and test sets? *Reference*: "Understanding Machine Learning: From Theory to Algorithms" by Shai Shalev-Shwartz and Shai Ben-David, Chapter 5 - Bias-Variance Decomposition

**Q15**. How can cross-validation be used to tune model complexity in terms of bias and variance? What are the limitations of cross-validation in addressing this tradeoff? *Reference*: "Introduction to Statistical Learning" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, Chapter 5 - Resampling Methods.

---

q-1)

For a given dataset, on increasing the complexity of a model how does the bias and variance change? Also give reasons for why that happens.

A). On increasing model complexity, bias decreases whereas variance increases.

B). On increasing model complexity, variance decreases whereas bias increases.

C). On increasing model complexity, bias and variance increase.

D). On increasing model complexity, bias and variance decrease.

**Correct Answer: A)**

**Explanation:**

On increasing model complexity, bias decreases whereas variance increases. Bias decreases as complex models have better fitting abilities i.e. they are able to fit well on the training data. However, this may lead to overfitting and so the model may not perform well on unseen data leading to higher variance.

**Rubric: 1 mark for correct option and explanation.**

**Q2.**

You are reviewing four papers submitted to a conference on machine learning for medical expert systems. All the four papers validate their superiority on a standard benchmarking cancer dataset, which has only 5% of positive cancer cases. Which of the experimental setting is acceptable to you?

paper i) We evaluated the performance of our model through a 5-fold cross validation process and report an accuracy of 93%.
paper ii) The area under the ROC curve on a single left out test set of our model is around 0.8, which is the highest among all the different approaches.
paper iii) We computed the average area under the ROC curve through 5-fold cross validation and found it to be around 0.75 – the highest among all the approaches.
paper iv) The accuracy on a single left out test set of our model is 95%, which is the highest among all the different approaches.

(A) paper i
(B) paper i and paper iv
(C) paper ii and paper iv
(D) paper iii

**Correct Answer: D)**

**Explanation:**

Paper iii: This paper reports the average AUC over a 5-fold cross-validation, which is a more robust and reliable evaluation method. It accounts for variability across different subsets of data, making the results more generalizable.

**Rubric: 1 mark for correct option. Reference: GATE 2024 Sample Paper**

**Q3.**

Which of the following functions can be related to or utilized within logistic regression for binary classification tasks? Select all that apply. Here x is the input.

**i)** $y = 1/(1 + e^{-x})$

ii) $y = ReLU(x)$

iii) $y = \tanh(x)$

iv) y=log(p/1-p)

(A) i
(B) i and ii
(C) i and iii
(D) i and iv

**Correct Answer: D)**

**Explanation:**

**Option A**: True. The function is the logistic function (or sigmoid function) and is the basis of logistic regression, used to model the probability of the positive class.

**Option B**: False. The ReLU function is an activation function used in deep learning models but is not related to logistic regression.

**Option C**: False. The hyperbolic tangent function is similar to the logistic function but maps input values to a range between -1 and 1. While useful in neural networks, it is not directly used in logistic regression.

**Option D**: True. The expression is the logit function, which is the inverse of the logistic function and is used in the formulation of logistic regression to express the linear relationship between the input features and the log-odds of the positive class.

**Rubric: 1 mark for correct option and explanation. Reference: Self**

**Q4:**

Is logistic regression a linear classifier? If yes, what decision boundary is used to divide the data into two classes (give the mathematical formulation)? Show how you got this decision boundary. If no, state the reason why.

**Solution:**

Yes, logistic regression is a linear classifier. The decision boundary for logistic regression is given by the following expression:

$$\sigma\left(w_0 + \sum_i w_i x_i\right) = 0.5$$

We need to show that this is a linear decision boundary.

$$\begin{aligned}
\frac{1}{1 + e^{-\left(w_0 + \sum_i w_i x_i\right)}} &= 0.5 \\
e^{-\left(w_0 + \sum_i w_i x_i\right)} &= 1 \\
w_0 + \sum_i w_i x_i &= 0
\end{aligned}$$

This is a linear decision boundary which shows that logistic regression is a linear classifier.