8BB020 Introduction to machine learning: Example questions

October 16, 2024

Question 1: Linear and logistic regression

1. Explain the key differences between linear regression and logistic regression in terms of the types of data they are typically used for. In your answer, specify the type model output (y) for both methods.
2. Provide a biomedical example for both methods. Describe a scenario involving biomedical data where they would be useful. Specify the input features and what the model predicts.
3. Consider the following formula:

$$p(y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

   Answer the following questions:

   - From which method is this formula?
   - What does the left-hand side of the equation represent in words?
   - What does the term inside the exponent ($\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$) represent?
   - What is the range of possible values for the left-hand side of the equation? Please provide it in the format [.;.] by filling in the dots.

Question 2: Regularization in linear models

Consider a standard linear regression model.

$$\begin{aligned} y &= \hat{y} + \epsilon \\ \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p \end{aligned}$$

where y is the measured output, $\hat{y}$ is the model prediction, ϵ represents the error (residuals) and $\beta_0, \ldots, \beta_p$ the model parameters to be estimated.

1. **Lasso and Ridge Regularization**

   - Explain the difference between Lasso (L1) and Ridge (L2) regularization in the context of linear regression. How do the types of penalties differ, and how do they affect the resulting model?
   - In the case of Lasso regularization, how is the regularization parameter λ incorporated into the cost function? Explain what happens to the coefficients as λ becomes very large, and link this to how increasing λ affects the model's bias and variance.

2. **Elastic Net regularization**

   - Elastic Net regularization is a combination of Lasso (L1) and Ridge (L2) regularization. Explain the advantages of using Elastic Net over Lasso or Ridge alone. In what situations would Elastic Net be more appropriate than using just Lasso or Ridge?
   - Mathematically, Elastic Net regularization is expressed as:

     $$RSS + \lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1 - \alpha)|\beta_j|)$$

     where RSS is the residual sum of squares.

     Describe the role of the parameter α in Elastic Net regularization. What happens when α is set to 0 or 1?

Question 3: Support vector classifier (SVC)

You are provided with the two-dimensional scatter plot below, showing data points classified into two classes, colored accordingly. The features are labeled as x_1 and x_2, and the separating hyperplane and corresponding margins are also shown.

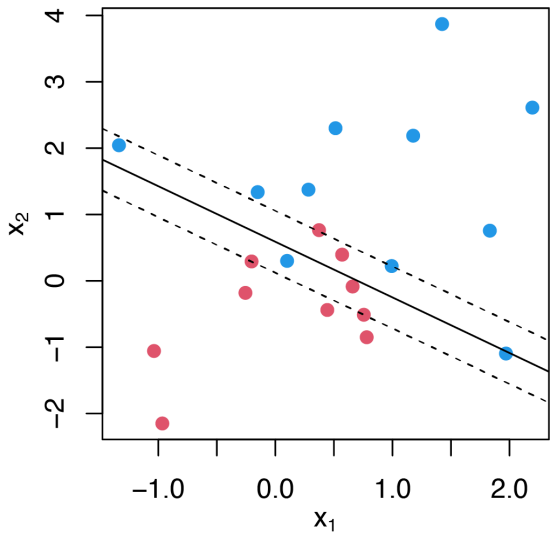

Figure 1: Support vector classifier trained on the shown data points, where observations belonging to class 1 are shown in blue and the one belonging to class -1 in red. The hyperplane is shown with a continuous line and the margins as dashed lines.

1. Mark the points that determine the orientation of the separating hyperplane (i.e., the support vectors) on the plot and briefly explain how support vectors are defined.

2. For each of the support vectors identified in part (1), report the value of the corresponding slack variable ϵ_i for $i = 1, \ldots, s$, where s is the number of support vectors.

   - $\epsilon_i = 0$
   - $0 < \epsilon_i \leq 1$
   - $\epsilon_i > 1$.

   Explain the meaning of each of the three categories.

Question 4: Tree-based methods

You are given a dataset and two models have been trained on it: - A single decision tree. - A bagging model (ensemble of decision trees using bootstrap aggregating).

1. Explain how a single decision tree is constructed. In your answer, describe the concept of recursive binary splitting and how the tree decides which feature and split point to use at each node.

2. Bagging combines multiple decision trees. Explain how bagging reduces the variance of the model. Why does bagging tend to outperform a single decision tree on a dataset with high variability or noise?

Question 5: Activation functions in neural networks

1. Which of the following activation functions would you use to prevent vanishing gradients? Give a short motivation for your choice.

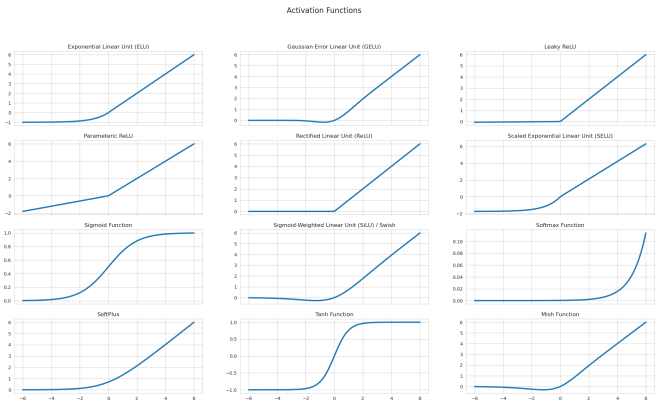

Figure 2: Different activation functions

2. After you solved the vanishing gradients some neurons stop outputting information. What is this phenomenon called, what is the cause and how can it be solved?

Question 6: Regularization of a neural network

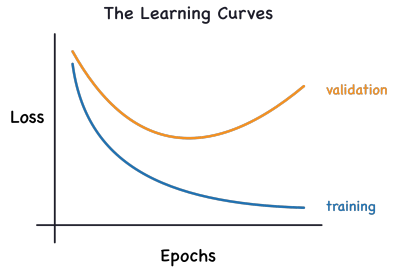

Figure 3: Performance of a neural network

The learning curve of a neural network is portrayed in the figure above. What would you do to elevate the performance of this particular model?

Question 7: Backpropagation

1. Why is backpropagation considered to be an efficient method for calculating gradients in deep neural networks?

2. Explain the relationship between backpropagation and gradient descent.

Given the following simple neural network of two layers:

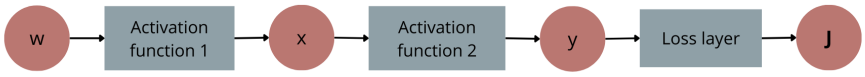

Figure 4: Computational graph

Activation function 1 is defined as sin(a)

Activation function 2 is defined as 3a+1

The loss can be calculated as $\mathbf{J} = (a - 5)^2$

3. Compute the gradient of the loss function **J** with respect to w using back-propagation. Initial input is equal to $w = 4$.

4. What is the concept of 'vanishing gradients,' and how does batch normalization help address this issue?

Question 8: Principal component analysis (PCA)

Consider the two scatter plots below, showing data points in two-dimensional space. In both plots, the axes represent two features, x_1 and x_2, measured for several data points.

1. Identify in which plot PCA would work better for dimensionality reduction, and explain why.

2. On the plot where PCA works better, draw the principal components (PC1 and PC2), clearly indicating their directions relative to the data.

3. In a few sentences, explain the role of the principal components in the dimensionality reduction process.

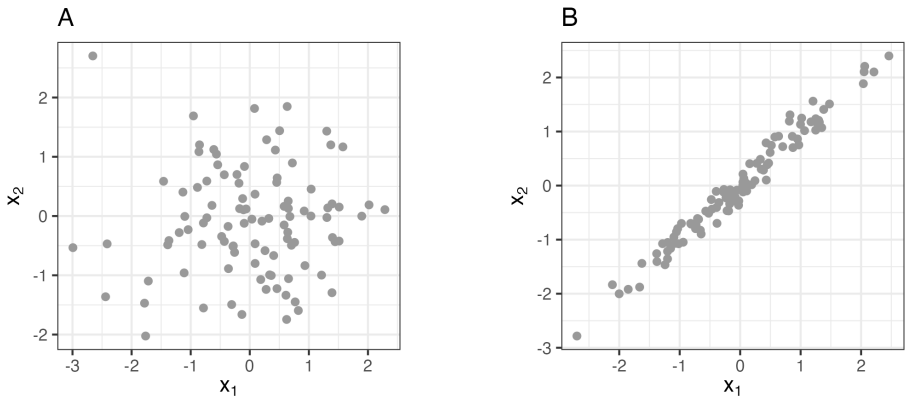

Figure 5: **A**. Data points exhibit little to no correlation between features x_1 and x_2. **B**. Data points are strongly correlation between features x_1 and x_2.

Question 9: Hierarchical clustering

You are given the following small dataset with two variables (x_1 and x_2) for 5 observations:

Observation	$\mathbf{x_1}$	$\mathbf{x_2}$
A	1.0	1.0
B	1.2	1.2
C	3.0	3.0
D	4.5	4.5
E	5.5	5.5

Which can be visualized as follows

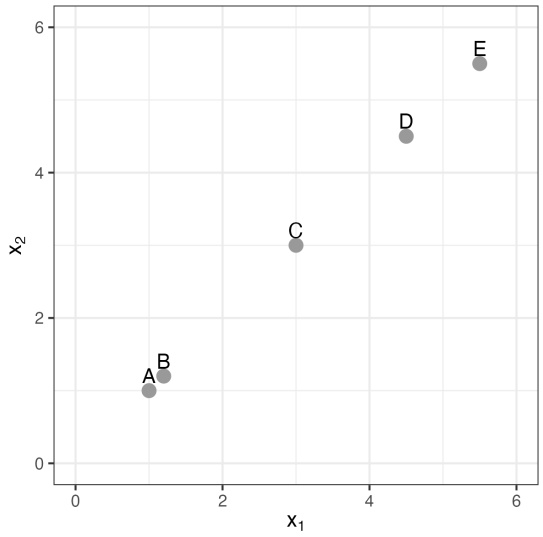

Figure 6: Plot of the data points from the table above.

1. Explain the differences between the single linkage, complete linkage, and average linkage methods in hierarchical clustering. In your explanation, describe how each method defines the distance between clusters.

2. Based on the dataset above, sketch how the dendrogram would change if you use:

   - Single linkage
   - Complete linkage

   Briefly explain how the choice of linkage affects the shape of the dendrogram.