Q1) Provide an example use case of where you will prefer precision over recall. Give justification [2 marks].

Answer:

Email Spam detection:This is one of the examples where Precision is more important than Recall.

Precision: This tells when you predict something positive, how many times they were actually positive. whereas,

Recall: This tells out of actual positive data, how many times you predicted correctly.

In the realm of spam email detection, it's acceptable for some spam emails (positives) to go undetected and not end up in the spam folder. However, it's crucial that legitimate emails (negatives) don't get wrongly classified as spam. Therefore, emphasizing Precision is key. Otherwise, there's a risk of missing important emails.

Q2) You are solving the binary classification task of classifying motifs. You design a CNN with a single output neuron. Let the output of this neuron be z. Now consider the following cases:

- **Case I - y = sigmoid(Leaky ReLU(z))**
- **Case II - y = Leaky ReLU(sigmoid(Leaky ReLU(z)))**

1. **y= sigmoid(Leaky ReLU(z))** [correct]

   Leaky Relu returns a small negative value for z<0, and z for z>0. So after passing through sigmoid all z<0 would map to <0.5 value. However, all the z>0 will map to >0.5 value. So it will give correct classification output.

   However the prediction probability might be changed.

2. **y = Leaky ReLU(sigmoid(Leaky ReLU(z)))** [correct]

   Output of sigmoid(Leaky ReLU(z)) is always positive (between 0 to 1). So passing through leaky relu does not have any effect, as it is passed unchanged.

Q3) Given model outputs of 90% and 99% accuracy for motif discovery classification.

- **Is it reasonable to conclude that the model with 99% accuracy (Model 2) is consistently superior to the one with 90% accuracy (Model 1)? [0.5 marks]**
- **Justify your choice [1.5 mark]**

No, it is not reasonable to conclude that the model with 99% accuracy (Model 2) is consistently superior to the one with 90% accuracy (Model 1) because accuracy alone is not a good evaluation metric for motif discovery classification (considering imbalance dataset). It's crucial to consider other metrics like precision, recall, and F1 score to get a more comprehensive understanding of a model's performance. Model 2 might have a higher

accuracy, but it's important to test its overall performance across different metrics before drawing conclusions about its superiority.

Q4) You are training a model where the training and test data are provided separately for gene expression classification. Surprisingly you found the number of samples in test dataset is 10 times more than training data. Also you observed there is no class imbalance in the training data, whereas the test data has high class imbalance. Do you think that you need to apply a strategy to handle data imbalance in the test set? Justify? If yes suggest two strategy for handling imbalance .[2 marks]

Data imbalance in the test set is not a problem. This is because the model is not trained on the test set, so the imbalance in the test set will not affect the model's performance. So no strategy is required to handle test data imbalance.

Q5) You are training a Deep learning model in two scenarios:
• Case I – Validation loss is significantly greater than training loss (almost zero).
• Case II - Validation loss converges to training loss but very high.
Case-1: Validation loss is significantly greater than training loss (almost zero).
This indicates that the model is overfitting. It can be handled by:

- Reducing the number of parameters or layers in the model.
- Implementing regularization techniques to penalize complex models

Case-2: Validation loss converges to training loss but very high.
This indicates that the model is underfitting. It can be handled by:

- Increasing the model complexity: This can be done by adding more features to the model, and increasing the number of layers in a neural network.
- Reduce regularization

Q6) Explain the process by which gradient descent updates weights on a parabolic loss curve to minimize the overall loss. Provide a diagram and a justification for your explanation. [2]

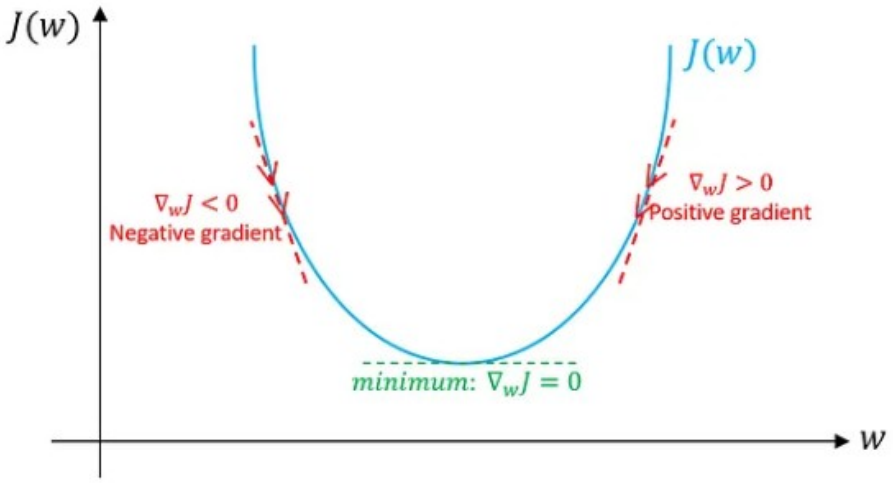


Figure 4: Gradient descent. An illustration of how gradient descent algorithm uses the first derivative of the loss function to follow downhill it's minimum.

[Image source: https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3]

Gradient descent is an optimization algorithm commonly used in machine learning to minimize a loss function. The loss function, often represented by a parabolic curve, measures the error between the predicted output of the model and the actual target values. Gradient descent aims to find the minimum point of this curve, which corresponds to the set of weights that minimizes the overall loss.

Imagine you're standing at the top of a hill and want to reach the bottom as quickly as possible. Gradient descent is like taking small steps downhill while always facing the steepest slope. The steepest slope indicates the direction of the gradient, which points towards the direction of the fastest decrease in loss. By repeatedly moving in the direction of the gradient, you eventually reach the bottom of the hill, representing the minimum loss.

In gradient descent, the weight updates are calculated using the following formula:

$$w(t+1) = w(t) - \alpha * \nabla E(w(t))$$

where:

- w(t) represents the weights at iteration t
- α is the learning rate, a small positive number that controls the step size
- ∇E(w(t)) is the gradient of the loss function E with respect to the weights w(t)

The process of gradient descent can be visualized as follows:

1. Initialize the weights to a random starting point.

2. Calculate the loss function and its gradient for the current weights.
3. Update the weights using the gradient descent formula.
4. Repeat steps 2 and 3 until the loss converges or a maximum number of iterations is reached.

Q7) In the regression case, the MSE loss is defined as the sum of the square error between the true and predicted labels. In the motif classification problem, we also found a loss function, which we try to optimize.

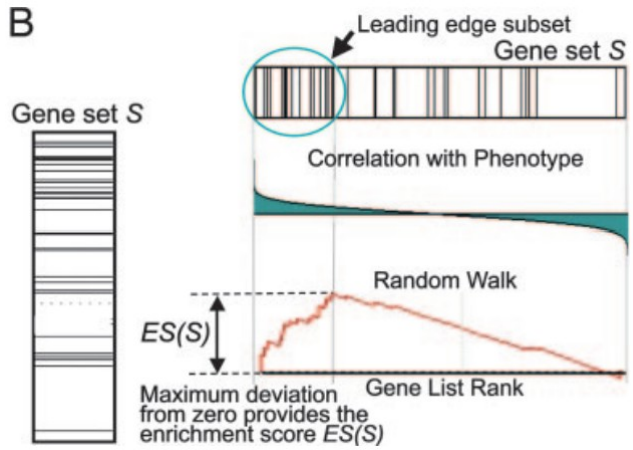
• What do you understand by loss function in case of motif classification (No mathematical formula required, only concept) [1 mark]

• Can you use MSE loss for motif classification? Justify [1 marks]

- In motif classification, the loss function serves as a measure of how well the model is performing in predicting the presence or absence of specific motifs in a sequence. It quantifies the difference between the predicted probabilities and the true labels assigned to the sequences.
- No, MSE is commonly used in regression problems where the output is continuous, but motif classification involves discrete categories (the presence or absence of a specific motif).

Q8. How does gene set enrichment analysis (GSEA) help us explore the impact of genes on biological pathways or functions
How do researchers use this method to understand the mechanisms behind diseases or biological processes? [3+2]

Sol- Gene set enrichment analysis helps us to explore the impact of genes on biological pathways or functions by quantifying the 'differential expression' of genes of a pathway. It is a computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states (e.g., phenotypes).
GSEA considers experiments with genome wide expression profiles from samples belonging to two classes, labeled 1 or 2. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric. Given an a priori defined set of genes S (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom. We expect that sets related to the phenotypic distinction will tend to show the latter distribution.



Steps are:
1: Calculation of an Enrichment Score
2: Estimation of Significance Level of ES
3: Adjustment for Multiple Hypothesis Testing

- Enrichment score ES=0
- Screen from top to bottom
  - If hit a gene in the pathway
    - $ES = ES + S_{hit}$
  - If hit a gene not in the pathway
    - $ES = ES - S_{unhit}$
  - $S_{hit} = \frac{|Correlation_g|^a}{\sum_{g \in pathway} |Correlation_g|^a}$
  - $S_{unhit} = \frac{1}{Number\ of\ gene\ not\ in\ the\ pathway}$

This method aids in understanding the mechanisms behind diseases or biological processes as:
In a typical experiment, mRNA expression profiles are generated for thousands of genes from a collection of samples belonging to one of two classes, for example, tumors that are sensitive vs. resistant to a drug. The genes can be ordered in a ranked list L, according to their differential expression between the classes. The challenge is to extract meaning from this list. A common approach involves focusing on a handful of genes at the top and bottom of L (i.e., those showing the largest difference) to discern telltale biological clues.

Q9. What does the term “enrichment score” means in GSEA and why is it important? How does it help to figure out if a specific group of genes is relevant to the experimental data? [3+2]

Sol –
In GSEA we calculate the “Enrichment score” that reflects the degree to which a priori defined set of genes S is overrepresented at the extremes (top or bottom) of the entire ranked list L. The score is calculated by walking down the list L, increasing a running-sum statistic

when we encounter a gene in S and decreasing it when we encounter genes not in S. The magnitude of the increment depends on the correlation of the gene with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk

- Enrichment score ES=0
- Screen from top to bottom
  - If hit a gene in the pathway
    - $ES = ES + S_{hit}$
  - If hit a gene not in the pathway
    - $ES = ES - S_{unhit}$
  - $S_{hit} = \frac{|Correlation_g|^a}{\sum_{g \in pathway} |Correlation_g|^a}$
  - $S_{unhit} = \frac{1}{Number\ of\ gene\ not\ in\ the\ pathway}$

Enrichment score helps to figure out if a specific gene set is relevant to experimental data. We estimate the statistical significance (nominal P value) of the ES. When an entire database of gene sets is evaluated, we adjust the estimated significance level to account for multiple hypothesis testing.

Estimating Significance. We assess the significance of an observed *ES* by comparing it with the set of scores ES_{NULL} computed with randomly assigned phenotypes.

1. Randomly assign the original phenotype labels to samples, reorder genes, and re-compute *ES*(*S*).
2. Repeat step 1 for 1,000 permutations, and create a histogram of the corresponding enrichment scores ES_{NULL}.
3. Estimate nominal *P* value for *S* from ES_{NULL} by using the positive or negative portion of the distribution corresponding to the sign of the observed *ES*(*S*).