Problem Set-5
Answers

2b) Consider that sequences marked with _n_ are truly non-coding and _c_ are truly coding. Conceptually, explain why some are misclassified.
Answer:

The misclassifications between coding and non-coding sequences from the output could be due to a variety of factors. These might include:

Sequence Complexity: Both coding and non-coding sequences can exhibit complex patterns. Coding sequences may contain regions that are not well conserved, which may resemble non-coding sequences. Conversely, non-coding sequences might have conserved elements due to their regulatory roles that resemble coding sequences.

Model Probabilities: The classification model uses probability matrices to evaluate whether a sequence is coding or non-coding. These matrices are constructed from known data and may not capture all biological variations. For example, rare codon usage or unusual evolutionary events that aren't well-represented in the training data can lead to misclassifications.

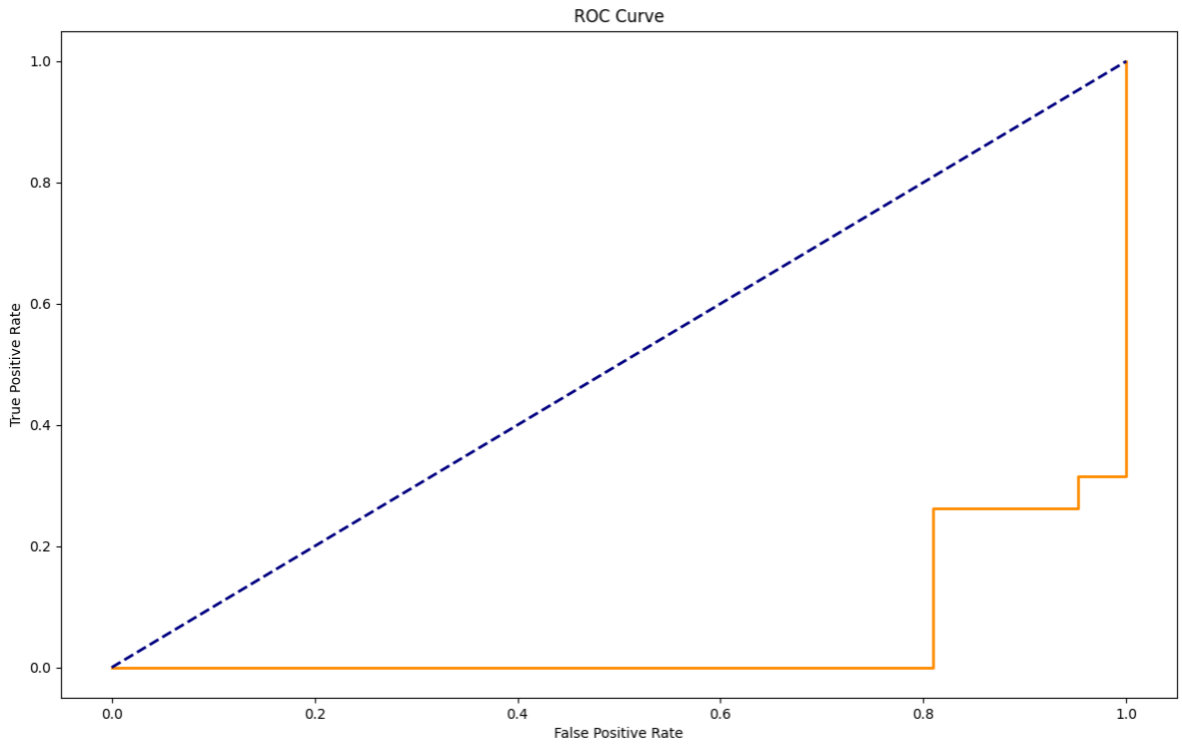
Evolutionary Pressure: Evolutionary pressures can make non-coding sequences appear similar to coding ones. For instance, some non-coding elements are under selective pressure to maintain certain sequences (like regulatory elements), which may accidentally mimic the patterns of coding sequences.

Statistical Noise: There is inherent noise in biological data that can lead to overlaps between the distributions of features characteristic of coding and non-coding sequences.

Pseudogenes: There are also genetic elements like pseudogenes that do not code for proteins but are derived from coding sequences. These elements can confuse the model because they have the hallmarks of coding sequences without actually being functional.

2c) Create an ROC curve with the results using the likelihood ratio as a threshold (what is the true and false positive rate for various threshold values). You are not required to code this. Based on the ROC curve, what cutoff would you use? Justify your answer – be sure to explain whether you might care more about specificity or sensitivity! (10 points)

Answer



I’m choosing to prioritize specificity over sensitivity, so that means I want to reduce the number of false positives, so that I can be more confident about my positive classifications. This would be important in situations where the cost of a false positive is high.

When looking at an ROC curve:

High specificity corresponds to a low False Positive Rate (FPR).
High sensitivity corresponds to a high True Positive Rate (TPR).
To maximize specificity, you want to choose a cutoff that is towards the bottom left of the ROC curve, where the FPR is as low as possible without sacrificing too much TPR.

From the ROC curve, this would typically mean you would look for a point on the curve that is closest to the bottom left corner but still provides an acceptable TPR. This point maximizes specificity and minimizes the FPR.

For example, if there's a point on the curve with an FPR of 0.1 and a TPR of 0.8, and another with an FPR of 0.2 and a TPR of 0.9, the first point would be the choice for higher specificity. The exact choice would still depend on how much TPR (sensitivity) you are willing to trade-off to gain in specificity.

In practice, this might mean you would set a higher threshold for deciding a sequence is coding, which would reduce the number of non-coding sequences incorrectly classified as coding (false positives), but may also increase the number of coding sequences missed (false negatives).

False positives are more detrimental than false negatives (e.g., in a legal context where falsely accusing someone has serious implications), so choosing specificity makes sense.

3a) Describe what the coding matrix will look like at time 2t. Show how you would calculate probability of TTT staying TTT after time 2t. No need to write out the full equation – you may use (...) – just show me enough to make it clear you follow the calculation by listing a few terms. (5 points)

Answer:

The coding matrix at time 2t will be the square of the coding matrix at time t. The entry for the probability of TTT staying TTT after time 2t will be the sum of the squares of the probabilities of TTT transitioning to each possible codon at time t and then back to TTT. If we denote the probability of transition from TTT to any codon X at time t as P(TTT -> X)_t, and the matrix entry at i, j as M(i, j), the probability of TTT staying TTT after time 2t can be expressed as

$$P(\text{TTT at } 2t \mid \text{TTT at } t) = M(\text{TTT}, \text{TTT})^2 + \sum_{x \neq \text{TTT}} (M(\text{TTT}, x) \times M(x, \text{TTT}))$$

The terms M(TTT, X) represent the probability of transitioning from TTT to any other codon X at time t, and M(X, TTT) represents the probability of transitioning from X back to TTT at time t.

3b) Make a prediction about your classification success if the sequences had longer to diverge. Explain how you would expect the ROC curve to change as t increases. (5 points)

Answer:

If sequences had a longer time to diverge, it is likely that the classification success would decrease. This is because, over time, mutations happen, and the sequences can drift, making the distinction between coding and non-coding sequences less clear.

How the ROC curve is expected to change as t increases:

The True Positive Rate (TPR) might decrease because coding sequences that have diverged for a longer time may acquire mutations that make them resemble non-coding sequences. As a result, the classifier may fail to recognize them as coding sequences.

The False Positive Rate (FPR) might increase because non-coding sequences might acquire mutations that make them resemble coding sequences, leading the classifier to incorrectly identify them as coding.