Classification

Thresholds and the Confusion Matrix

A logistic regression model outputs probabilities between 0 and 1, representing the likelihood an email is spam. To use these predictions, a classification threshold is set (e.g., 0.5) to convert probabilities into labels like "spam" or "not spam." Emails scoring above the threshold are marked as spam. However, choosing the right threshold is crucial—especially when classes are imbalanced or when the cost of false positives (e.g., misfiling important emails) is high.

Confusion Matrix

There are four possible outcomes for each output from a binary classifier. For the spam classifier example, if you lay out the ground truth as columns and the model's prediction as rows, the following table, called a **confusion matrix**, is the result

	Actual positive	Actual negative
Predicted positive	**True positive (TP)**: A spam email correctly classified as a spam email. These are the spam messages automatically sent to the spam folder.	**False positive (FP)**: A not-spam email misclassified as spam. These are the legitimate emails that wind up in the spam folder.
Predicted negative	**False negative (FN)**: A spam email misclassified as not-spam. These are spam emails that aren't caught by the spam filter and make their way into the inbox.	**True negative (TN)**: A not-spam email correctly classified as not-spam. These are the legitimate emails that are sent directly to the inbox.

Accuracy, Recall, Precision and Related Metrics

Accuracy is the proportion of all classifications that were correct, whether positive or negative. It is mathematically defined as:

$$Accuracy = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

In the spam classification example, accuracy measures the *fraction of all emails correctly classified.*

A perfect model would have zero false positives and zero false negatives and therefore an accuracy of 1.0, or 100%.

The **true positive rate (TPR)**, or the proportion of all actual positives that were classified correctly as positives, is also known as **recall**.

Recall is mathematically defined as:

$$Recall = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

The **false positive rate (FPR)** is the proportion of all actual negatives that were classified *incorrectly* as positives, also known as the **probability of false alarm.** It is mathematically defined as:

$$FPR = \frac{\text{incorrectly classified actual negatives}}{\text{all actual negatives}} = \frac{FP}{FP + TN}$$

Precision is the proportion of all the model's positive classifications that are actually positive. It is mathematically defined as:

$$Precision = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

ROC and AUC

Receiver-operating characteristic curve (ROC)

The **ROC curve** is a visual representation of model performance across all thresholds. The long version of the name, receiver operating characteristic, is a holdover from WWII radar detection.

The ROC curve is drawn by calculating the true positive rate (TPR) and false positive rate (FPR) at every possible threshold (in practice, at selected intervals), then graphing TPR over FPR.

A perfect model, which at some threshold has a TPR of 1.0 and a FPR of 0.0, can be represented by either a point at (0, 1) if all other thresholds are ignored

Area under the curve (AUC)

The **area under the ROC curve (AUC)** represents the probability that the model, if given a randomly chosen positive and negative example, will rank the positive higher than the negative.

The perfect model above, containing a square with sides of length 1, has an area under the curve (AUC) of 1.0. This means there is a 100% probability that the model will correctly rank a randomly chosen positive example higher than a randomly chosen negative example. In other words, looking at the spread of data points below, AUC gives the probability that the model will place a randomly chosen square to the right of a randomly chosen circle, independent of where the threshold is set.

Prediction Bias

Prediction bias is the difference between the mean of a model's **predictions** and the mean of **ground-truth** labels in the data. A model trained on a dataset where 5% of the emails are spam should predict, on average, that 5% of the emails it classifies are spam. In other words, the mean of the labels in the ground-truth dataset is 0.05, and the mean of the model's predictions should also be 0.05. If this is the case, the model has zero prediction bias. Of course, the model might still have other problems.

Prediction bias can be caused by:

- Biases or noise in the data, including biased sampling for the training set
- Too-strong regularization, meaning that the model was oversimplified and lost some necessary complexity
- Bugs in the model training pipeline
- The set of features provided to the model being insufficient for the task

Completion