Class 6 Review Notes

AI & Machine Learning
Fall 2023 - Laurie Ye

6 Assessing the Performance of Classifiers

6.1 Confusion Matrix

The confusion matrix is a table that shows the true categories vs. the predicted categories. The true categories usually go on the left. The predicted categories usually go on the top.

6.2 Basic Measures

- Accuracy = % correct classifications.
- Error Rate = % incorrect classifications=1-Accuracy.
- Sensitivity for a category is defined as the probability of correctly classifying observations from that category:
$$P(\text{Classified as } X|\text{True } X)$$
Sensitivity is also known as recall.
- Precision is the probability that an item classified as X truly is X:
$$P(\text{True } X|\text{Classified as } X)$$

	Classified as Spam (Test +)	**Classified as Not Spam (Test -)**
Is Spam (Is +)	True Positive (TP)	False Negative (FN)
Not Spam (Is -)	False Positive (FP)	True Negative (TN)

Table 1: Confusion Matrix for Email Classification

6.3 Connection to Hypothesis Testing

- Positive test result means null hypothesis (H0) is not true. That is, HA is true.
- Negative test result means null hypothesis (H0) is true.
- Testing positive means reject H0.
- Testing negative means do not reject H0.

	Reject Test	**Fail to Reject Test**
H0 is False	Correct	Type II Error
H0 is True	Type I Error	Correct

Table 2: Outcomes of Statistical Tests

	Reject Test	**Fail to Reject Test**
H0 is False	Probability is $1 - \beta$ (the power)	$P(\text{Type II Error}) = \beta$
H0 is True	$P(\text{Type I Error}) = \alpha$ (the significance level)	Probability is $1 - \alpha$

Table 3: Statistical Test Outcomes

Note: β (the power) depends on how far away the true parameter value is from the null hypothesis.

6.4 Binary Classifier Performance

- Accuracy: $P(\text{True}+) + P(\text{True}-)$
- Error Rate: $P(\text{False}+) + P(\text{True}-)$
- Sensitivity or Recall: $P(\text{Classified} + |\text{Is}+)$
- Specificity: $P(\text{Classified} - |\text{Is}-)$
- Precision: $P(\text{Is} + |\text{Classified}+)$

6.5 F1 Score

- The F1 score tries to balance recall and precision:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$= 2 \times \frac{P(\text{True} + |\text{Classified}+) \times P(\text{Classified} + |\text{True}+)}{P(\text{True} + |\text{Classified}+) + P(\text{Classified} + |\text{True}+)}$$

Note: $0 \leq F_1 \leq 1$

6.6 New Measure: ROC Curve

- The binary classifier has some parameter that adjusts its sensitivity.
- Main Idea of ROC Curve: Plot the probability of true positives against the probability of false positives.

- Plot: $P(\text{Classified} + |\text{True}+)$ vs. $P(\text{Classified} + |\text{True}-)$
- Plot the sensitivity vs. the $1-$ specificity.

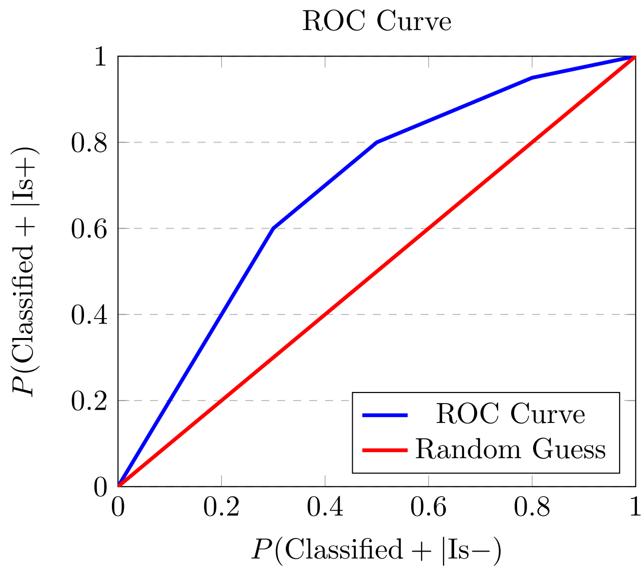

Figure 1: ROC Curve demonstrating the performance of a binary classifier

1. The more the (blue) curve stretches up and to the left, the better.
2. The 45° line corresponds to classifying randomly with $p = P(\text{Classified } + | \text{True } +) = P(\text{Classified } + | \text{True } -)$, the probability of classifying as '+'.

6.7 AUC

- AUC = Area Under the Curve
- AUC is bounded between 0 and 1.
- The bigger the AUC the better; that is, the closer to 1 the AUC is the better.
- Random classifier corresponds to an AUC of 0.5.
- So for any sensible procedure, $0.5 < \text{AUC} \leq 1$

6.8 Other Performance Measures

Some classification methods (most notably regression trees) find regions (in the x-space) and estimate the probability of each class in this region by the sample proportion of each class.

Gini Index

The Gini index is a measure of how "pure" a region is. Specifically, the Gini index for a region m with K classes is

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Small values indicate a high degree of node purity

Entropy

Entropy is a measure similar to the Gini index. The entropy for region m for K classes is

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Note: $D_m \geq 0$ and $\lim_{p \to 0} p \log p = 0$

The entropy will be close to 0 when the node has high "purity."

Log Likelihood

The log likelihood is also often used as the loss function in machine learning problems for classification. Specifically, for n observations where $\hat{p}_{ik}$ is the probability predicted by the model that the i'th observation belongs to the k'th class,

$$\text{log likelihood} = \sum_{i=1}^{n} \sum_{k=1}^{K} Y_{ik} \log \hat{p}_{ik}$$

where $Y_{ik} = 1$ if the i'th observation belongs to the k'th class and 0 otherwise.

Cross-Entropy

- The cross-entropy of a distribution q relative to a distribution p is defined to be:

$$H(p, q) = -E_p[\log q]$$

- For discrete distributions (the case we primarily care about), this is

$$H(p, q) = -\sum_{k=1}^{m} p(v_k) \log q(v_k)$$

Cross-Entropy (cont.)

- Note: Both distributions are assumed to take on the discrete values.
- If Y is a one-hot encoding of the categories defined by $v_1, v_2, \ldots, v_m$, then

$$E(Y_k) = p_k$$

- If we were to sample Y (the one-hot encoding) repeatedly (say n times), then we would obtain a data matrix of 0's and 1's:

$$[Y_{ik}]_{n \times m}$$

- The column averages of this matrix would estimate the $p(v_k)$'s.
- Thus,

$$\hat{H}(p,q) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} Y_{ik} \log q(v_k)$$

  is an estimate of the cross entropy $H(p,q)$.
- The double summation is the log likelihood.
- So, the cross entropy is proportional to the negative of the log likelihood.
- In MLE, the distributions being "crossed" are the empirical distribution of the data and the distribution predicted by the model. - So, maximum likelihood estimation is equivalent to finding the parameter estimates that minimize the cross-entropy.

6.9 Reference

Goizueta Business School-Emory University: Professor George S. Easton