Intro Deep Learning: Homework 2

Daniel Novikov

February 24th, 2022

Problem 1

Accuracy is probably the most commonly used metric in classification problems. Which statement is True?

Answer: B Accuracy is non-differentiable and cannot be used for direct optimization via gradient descent.

Problem 2

You are working as a Machine Learning Engineer in Metflix Inc. You are building a model to classify users who watch a lot of movies in Ultiverse. What metrics will you choose to evaluate your model?

Answer: As this is an unsupervised learning problem, we don't have a true label to calculate many of the common error metrics we normally use. However, if we can identify a few users manually that like the ultiverse and also some that dislike it, then we can start off by treating this as a clustering problem. Let's suppose we pick a diverse subset of users that is representative across many genre preferences. And let's also pick a few that we know like Ultiverse and those that we know don't. We cluster all users by their distance to the nearest representative user. Now for each cluster that contains a manually selected user with a known opinion on Ultiverse, we color that cluster as likes (or dislikes) ultiverse, respectively. Let's say the 'weight' of each cluster is the proportion of users in the cluster that like ultiverse, in the range [0,1]. For all the other clusters that don't have one of the known-opinion users, we leave their like/dislike status unknown = 0.5. Now, we test by recommending ultiverse to a few user from each cluster. Then we can compute the loss as the difference between the proportion of users that clicked ultiverse and the weight given by the cluster, and update the weight of the cluster accordingly.

Problem 3

Which method is used involved in numerical optimization of an appropriate selection of model criterion? How do you define the error of such estimator?

Answer: K-fold cross validation gives us a way to test the performance of a model on a limited set of data. It involves holding out one of the k folds as a testing set during each iteration. **Grid Search Cross Validation** runs K-fold cross validation on different choices of parameters or models. When choosing model criteria and parameter choices, trying each configuration once, by holding out the same test set, risks the test set having some unique characteristics that do not generalize to the real data distribution, which leads to misleading scores for each model. K-fold cross validation reduces this risk by going through the entire dataset and holding out each fold throughout iteration. Using Grid Search Cross Validation gives us K scores for each choice of model and parameters, and these scores can be averaged to give a more robust performance measure.

Problem 4

Attached separately in Github.

Problem 5

In the above figure we have ROC curves for two classifiers (A and B) which have equal areas under the curve (AUC).

- (a) Which classifier is better among these two? (write your detailed thinking)
- (b) Describe a situation in which undoubtedly classifier A should be preferred.
- (c) Describe a situation in which undoubtedly classifier B should be preferred.
- (d) Which factors determine the area under the curve

a) It depends on what we are classifying and the cost of false positives. If we want to maximize true positives, then classifier B is better. It gets nearly 100% TPR at FPR = 0.5. While A can't get near 100% without always classifying positive. If we want to minimize false positives, then A is better, as, for only FPR = 0.15, it catches over 60% of true positive cases, while B only catches 40%

b) If we are hiring job applicants, hiring a false positive can be much more costly than missing out on good candidates.In this case classifier A is better.

c) If someone is being tested for cancer, we would want to catch every positive case even if some people get false positives. In this case classifier B is better.

d) The Area under the curve is a quality measure of a classifier, and the top left of the range represents a prefect classifier, one that correctly predicts positive for all positive samples, and only those samples. Since the curve is a

function, it won't have any weird diagonal spikes that would violate the vertical line test. Therefore the closer the curve approaches the top left of the graph, the more area it will have. The area is determined by the proportion of positive samples the model can classify positive. If a model cannot classify more than 20% of positive samples correctly except at very high false positive rates, then it will have a low area under the curve. Conversely, if it can classify all positive samples with no false positives, then the area under the curve will be 1.

Problem 6

Technically we need to compute the gradient with respect to Wi , the linear transform (or parameter) matrix (tensor) for layer i. Yet, we are computing gradient of the loss with respect to the input x. How come?

Answer We compute the derivative with respect to the inputs because it tells us the degree to which each input feature contributes to the loss, informing which weights need to be updated. For example, if feature 3 contributes more to the loss than feature 2, it will have a higher gradient. Correspondingly, component 3 of the weight tensor will receive a larger update than component 2.

Problem 7

Compare the following metrics and explain which one is better. Image shows Sensitivity vs Specificity plots for 3 curves. Curve 1 has AUC=1, Curve 2 has AUC = 0.8, Curve 3 has AUC=0.5.

Answer a Curve 3 with AUC=0.5 is best.
Sensitivity (TPR) is number of positive predictions/ number of positive samples
Specificity (TNR) is number of negative predictions/ number of negative samples

Curve 1 with AUC=1 has high sensitivity and low specificity, which means the model will accurately classify all positive samples but misclassify all negative samples. Thus this classifier predicts positive for all samples

Curve 2, having an AUC = 0.8 means the model is more discriminative but still very eager to assign the positive class.

Curve 3, with AUC=0.5 is the optimal model, which accurately classifies all positive and negative samples.

Answer b The x axis in both the sens vs spec and TPR vs FPR planes relates to how the model handles negative samples. TNR and FPR are complementary with respect to the negative samples. That's why in the TNR vs FPR plane, a curve like curve 1 would be best, and a curve like curve 3 would be worst, but in the sens vs spec plane, it is the opposite.