Data C182 Designing, Visualizing & Understanding DNN
Fall 2024 Eric Kim, Naveen Ashish

Discussion 10

This discussion covers evaluation metrics.

1. Classification Question Statement

Let's consider a standard binary classification problem. Let's say you've trained a neural network to take in pictures of cats and dogs and perform binary classification to say which of those two kinds of animals they contain. Let's also say that their two outputs look like a probability distribution: the numbers corresponding to the two categories are non-negative and sum to one.

Problem: Review - Neural Network Architectures

What would an appropriate neural network architecture be for this problem, i.e., it takes in images and outputs something that looks like a probability distribution over the categories "cat" and "dog"?

Now, let's say you pass an image of a dog or cat through the network. It gives you two scores, one corresponding to each category. Based on this, how do you choose which category you think the image came from?

2. Calibration

A pretty natural idea is to pick a **decision threshold** value: if the score for dog exceeds that amount, then you say the image contains a dog. Otherwise, you say that the image contains a cat. But then the question remains: how do we pick that threshold value?

Let's consider a simple case first: when your model outputs are **calibrated**. This means that, when your model assigns a probability score to a category (e.g., "This image is 60% likely to contain a dog and 40% likely to contain a cat"), then those probabilities are correct: the image *actually* has a 60% of containing a dog.

In practice, to measure whether your model is calibrated or not, you pass in a bunch of images and put them into bins based on the model's assigned probability scores: e.g., bin all images that have $0-10\%$ probability score for dog, then $10-20\%$, etc. If your model is actually calibrated, the images in that first bin should actually be dogs with the probability assigned to them by the model (e.g., $0-10\%$ of the images in the first bin should be dogs).

Problem: Thresholding Calibrated Networks

In the case where your model is calibrated, what should your threshold be?

A lot of the time the "standard" threshold is not necessarily the best one to use though. For example, suppose that in the dataset of dog and cat pictures that you want to classify, your model assigns pictures of dogs only 40% probability of being dogs, while it assigns cats 10% of being dogs (maybe because the dogs in this dataset look particularly like cats). In this case, your model isn't calibrated (or it might be calibrated on a different distribution of dog and cat pictures), and using the 0.5 threshold will classify none of the dogs correctly. How else can we choose the score then?

One idea is to find some **evaluation metric** that measures how well your model is classifying stuff, akin to the loss you used for training, then pick a threshold that empirically maximizes that score. We will discuss two metrics that we could care about.

The first is **precision**, defined as the portion of datapoints correctly classified a particular way divided by the total number of datapoints classified that way. In this case, that would be the number of dog pictures correctly classified as dogs divided by the total number of pictures that the model classified as dogs *correctly or incorrectly*.

Problem: Precision Intuition

When would you want to maximize precision? What would a safe strategy for maximizing precision look like?

The second is **recall**, defined as the portion of datapoints correctly classified a particular way divided by the total number of instances of that class. In this case, that would be the number of dog pictures correctly classified as dogs divided by the total number of pictures of dogs in your dataset.

Problem: Recall Intuition

When would you want to maximize recall? What would a safe strategy for maximizing precision look like?

Problem: Precision and Recall Calculation

Suppose you have a dataset of 10 dogs and 12 cats. Your model classifies 8 images as dogs, 7 of which were actually dogs. What are the precision and recall of your model?

Usually though, you'd want some in-between metric that balances between these two. The metric balancing the two is **F1 score**, the harmonic mean between the precision and recall:

$$F_1 = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} \tag{1}$$

Each of these metrics is computed for a given threshold (or whatever other class decision strategy you have). So, to find a threshold that maximizes the metric, simply test out a wide range of thresholds and see which one achieves the highest. A lot of the time, people visualize this by plotting a **precision-recall curve**: a 2D plot with precision on one axis and recall on the other (both from 0 to 1), with all points on the curve being empirical precision and recall from some threshold value you tried out.

Problem: Which Metric?

Suppose your classification task is to figure out whether or not medical patients have some disease based on some measured biomedical information. Consider two cases.
1. The disease is not very serious, though patients with the disease would still definitely prefer to be cured of it. If the model says the patient has the disease, they will undergo an expensive treatment (though patients with the disease do think the expense is worthwhile).
2. The disease is potentially serious, but the medication is both inexpensive (maybe it's covered by most insurances or something!) and non-harmful to patients without the disease.
In each case, which metric should you intuitively try to maximize?