Evaluation

Sharad Chitlangia

05-11-2021

Evaluation

When we evaluate models, we will need a loss function and our goal is to minimize the expected loss value.

The reason for it to be *expected* loss is because our loss is always going to be computed over a sample of data points. So in the probabilitic sense, the loss is associated with an actual distribution.

Suppose we're given $(x_1, y_1), ..., (x_n, y_n)$ and a new point $(x_i, ?)$. We want to find the value of $\hat{y}$ that minimizes $L(y, \hat{y})$ for the point

Problem: We do not know y, so we cannot really figure out if something is minimizing $L(y, \hat{y})$.

Suppose we're given data $D : (x_1, y_1), ..., (x_n, y_n)$ which is drawn from a distribution $D_{x,y}$. We want to predict $\hat{y}$ so as to minimize $L(y, \hat{y})$ for any (x, y) drawn from the distribution $D_{x,y}$.

Problem: It is similar to the previous problem but worse. Here we don't know the values of (x, y), so how de minimize the loss?

Solution: In both the only solution is to treat X and Y as random variables and minimize expected loss.

Case 1: Given x i.e., X=x

$$\mathbb{E}[L(y, \hat{y})] = \sum_{y} L(y, \hat{y}) P(Y = y | X = x)$$

This is called as conditional risk. If we assume that we have $0 - 1$ loss function:

$$\mathbb{E}[L(y, \hat{y})] = \sum_{y \neq \hat{y}} P(Y = y | X = x)$$

which further reduces to:

$$\mathbb{E}[L(y, \hat{y})] = 1 - P(Y = \hat{y} | X = x)$$

So minimizing the expected loss is similar to maximizing $P(Y = \hat{y} | X = x)$. Note that this is the right decision to take because we are using a $0 - 1$ loss function.

Assume a squared loss function: $L(y, \hat{y}) = (y - \hat{y})^2$

$$\mathbb{E}[L(y, \hat{y})] = \int (y - \hat{y})^2 f(y|x) dy$$

To minimize this predict $\hat{y} = \mathbb{E}[y|x]$. Proven by setting $\frac{\partial \mathbb{E}[L(y, \hat{y})]}{\partial \hat{y}} = 0$

Case 2: Both x and y are unknown

$$\mathbb{E}[L(y,\hat{y})] = \sum_{x,y} L(y,\hat{y})P(X=x,Y=y)$$

We can make this a bit more explicit by specifying the prediction $\hat{y}$ to be a function of x.

$$\mathbb{E}[L(y,\hat{y})] = \sum_{x,y} L(y,f(X=x))P(X=x,Y=y)$$

Note that $f(x)$ here is conditional on the given dataset d which can be thought of as an instance of the random variable D.

$\mathbb{E}[L(y,\hat{y})|D=d]$ is called the generalization error.

We can estimate the generalisation error using a train-test split. But the problem then becomes that we are not really estimation the generalisation error on d then. It is rather a subset of d. There are three possible solutions to this: 1. Request more data so that we can train on d and test on the extra data. 2. Generate more data ourselves using a generative model e.g. $P(X,Y)$. 3. We just provide the training estimates i.e., we train and test on the same set.

Solution: To avoid the problems associated with calculating the generalisation error given a dataset d, we can treat d as an instance of a random variable D and calculate the expected generalisation error (e_m) rather than the generalisation error:

$$EGE(e_M) = \mathbb{E}(e_{M,d}) = \sum_{x,y} L(y, f_{M,d}(x))P(x,y|D=d)P(D=d)$$

The advantage here is that the error is not calculated for one instance of the dataset but rather considering all possible datasets at once.

So now we can repeatedly do train-test splits to estimate e_M without facing any of the earlier problems. Here, the dataset we have $D=d$ acts as a proxy for the population. By this we mean that we assure that sampling from d is equivalent to sampling from D.

Now in the case that we had estimated $e_{M,d}$ using the entire training dataset, then our estimate for $e_{M,d}$ will be *optimistic*, i.e., given new data $d\prime$:

$$e_{M,d\prime} = e_{M,d} + \epsilon$$

where ϵ is the training error or *optimism*.

Evaluating two models for the same problem

We cannot simply compare accuracies of two different models, as the results might be sampling dependent. We need to find a way to check if two models are doing the same thing and whether the differences in performances of the two models is meaningful.

.	Correct	Incorrect
Correct	n_1	n_2
Incorrect	n_3	n_4

If both the models are doing the same thing, then the points at which they differ is just some random noise.

$$
\begin{aligned}
Expected\ Count\ of\ A = 0, B = 1 &= \frac{n_2 + n_3}{2} \\
Expected\ Count\ of\ B = 0, A = 1 &= \frac{n_2 + n_3}{2}
\end{aligned}
$$

We can use χ^2 test to check for the independence of the two variables above. If we find that the two models are not doing the same thing, try and identify where one of the models does better so we can use that information to improve.

Evaluating two models on many problems

Dataset	M_1	M_2
1	A_{11}	A_{21}
2	A_{12}	A_{22}
.	.	.
.	.	.
n	A_{1n}	A_{2n}

Suppose we label the datasets in which M_1 is better as **heads** and datasets in which M_2 is better as **tails**.

If M_1 and M_2 are doing the same thing, then the probability of heads and tails should roughly be equal. We can now make a probabilitic estimate whether M_1 and M_2 are doing the same thing.

This, ofcourse, is not taking into account the magnitude of difference in performance for each dataset. A solution is to use the Wilcoxon signed rank test.

Evaluating many models on many problems

ROC Curve True Positive Rate (TPR):

$$
\begin{aligned}
TPR &= \frac{TP}{TP + FN} \\
TPR &= 1 \ - \ FNR
\end{aligned}
$$

False Positive Rate (FPR)

$$FPR = \frac{FP}{FP + TN} = 1 - TNR$$

Each model will have a point on the ROC Curve for a given dataset.

So far we have mostly assumed that a false positive and a false negative are equally penalized. In reality, this might not be the case.

$$Loss = P(+)P(-|+)C(-|+) + P(-)P(+|-)C(+|-)$$

$$Loss = P(+)C(-|+)(1 - TPR) + P(-)C(+|-)FPR$$

This loss is a linear function of FPR and TPR which define the space of the ROC curve.