Cross Validated

Model fitting vs minimizing expected risk

Asked 2 years, 11 months ago Active yesterday Viewed 195 times

9

3

The bounty expires in 20 hours. Answers to this question are eligible for a +50 reputation bounty. user27886 wants to **draw more attention** to this question.

I'm confused about the mechanics of model fitting vs minimizing risk in decision theory. There's numerous resources online, but I can't seem to find a straight answer regarding what I'm confused about.

Model fitting (via e.g. maximum log-likelihood):

Suppose I have some data pairs $\{(x_1, y_1), \dots, (x_N, y_N)\}$ and I want to come up with a parametric probability density modelling target y given x:

$$p(y|x;\theta)$$

which I use to estimate the true conditional distribution of the data, say $p_{\text{true}}(y|x)$. I can do so via some procedure to e.g. maximize log likelihood:

$$\max_\theta \sum_i \log p(y_i|x_i;\theta)$$

Then on future unseen data for x, we can give e.g. confidence intervals for its corresponding y given x, or just report $y_{\text{guess}} = y_{\text{mode}} = \arg\max_y p(y|x;\theta)$. y and x can both be continuous and/or discrete.

Decision theory:

A problem comes when we want a point estimate of y and the optimal point for an application is not captured purely by which is most frequent or expected, i.e. we need to do better than picking the modal

$$y_{\text{guess}} = \arg\max_y p(y|x;\theta)$$

or expected value

$$\mu_{y|x} = \mathbb{E}_{p(y|x;\theta)}[Y|x]$$

for said particular application.

So suppose I fit a model using maximum likelihood, then I want to make point predictions. Since I must pick a single point, I can predict a new point which minimizes expected cost; I choose the y_{guess} with the lowest avg cost along all y:

$$y_{\text{guess}} = \arg\min_{y} \int_{y'} L(y, y')p(y'|x;\theta)dy'$$
$$= \arg\min_{y} \mathbb{E}_{p(y'|x;\theta)}\Big[L(y, y')\Big]$$

This is the degree to which I understand decision theory. It's a step that you take *after* one has fit their model to pick point estimates of y and one has a loss function $L(y, y')$, when your model gives an entire *distribution* of y, but we need a point estimate, y_{guess}.

Questions:

- If the loss $L(y_{\text{guess}}, y')$ is what we actually care about minimizing in the pursuit of obtaining *point* estimates, then why not do the following fitting procedure instead of maximum likelihood:

$$\min_{\theta} \sum_{i} \int_{y'} L(y_i, y')p(y'|x_i;\theta)dy'$$

that is, minimize the expected loss under the parametric model $p(y|x;\theta)$? My current understanding is approach is called "Expected Risk Minimization" and this is done in practice sometimes, but the parametric model in this case would lose the interpretation as the approximation to the true distribution $p_{\text{true}}(y|x)$. Is my understanding correct? Are there any problems with doing this?

maximum-likelihood | inference | loss-functions | decision-theory | risk

Share Cite Edit Follow Flag

edited yesterday
Richard Hardy
50.4k 10 89 209

asked Oct 10 '18 at 15:13
user27886
525 4 13

not sure i understand your setup. After training a model you have only point estimation. i.e. for given weights the model will output only one number as prediction and not distribution. can you clarify? – ofer-a Sep 21 at 7:48

$p(y|x;\theta)$ is the conditional distribution or conditional density for y given an x and parameter value (or values) θ. Once we've fit a θ using maximum likelihood on a training set, we can simulate (draw samples) of y given unseen x values i.e. $y \sim p(y|x;\theta)$. So Y is a random variable with conditional distribution $p(y|x;\theta)$ and y represents a particular value or sample for Y. I hope my notation is clear enough. Let me know if something is still wrong/confusing. – user27886 Sep 21 at 20:12

1 Answer

Active | Oldest | Votes

2

One of the problems with "Expected Risk Mimimization" is that the distribution, $p(y'|x_i;\phi)$, is unknown, and thus it is not clear how to minimize it.

2

You could theoretically argue, that we can try to find this distribution, during the minimization process, but as you said, in this case it will not approximate the true distribution, because the model incentive is minimize the loss and not finding the probability distribution. For example, It can end up with putting the mass probability on the a single point that minimizes the loss, instead of finding the full distribution. If all we care about is minimizing the loss (as opposed to approximating the true distribution), we could minimize the loss directly e.g. minimize the MSE. i.e.

$\min_\theta \sum_i MSE(y_i, y_i^{'})$

Moreover, this is just theoretically. Practically, true distribution is intractable and cannot be computed. This is also true for the maximum likelihood optimization.

Usually in machine learning, in order to solve the intractability problem, we add assumptions regarding the output distribution. We tend to pick "easy to work with" distribution and not complex one. For example distributions that are easily described by finite numbers like μ, σ. A popular choice for such distribution is the normal distribution. i.e. we assume that the the output is normally distributed with the label as mean plus some variance around it.

It can be proved that, assuming normal distribution, maximizing the log likelihood is equivalent to minimizing the MSE. So in the practical setup, in which you assume normal distribution, both tasks are the same. i.e. in one shot, we are minimizing the loss and maximizing the likelihood. See more details here

Share Cite Edit Follow Flag edited Sep 22 at 11:24 answered Sep 22 at 11:18

ofer-a
738 3 9

I generally like to use Bayesian parameter estimation in conjunction with optimization techniques. I find that obtaining a posterior chain over $p(y|x)$ is more robust than point estimates (whether that be a single y or (y_μ, y_σ). So when it comes time to optimize, your loss function can iteratively draw from the posterior chain, giving you a more robust sense of how the optimal y compares to $p(y|x)$. – jbuddy_13 5 hours ago