6.7900 Fall 2024: Lecture Notes 7

1 Model Evaluation

In previous lectures, we talked about how to quantify uncertainty in regressions. In this lecture, we will take a step back and consider how to evaluate models for supervised learning in general. Let us start with a simple question: is it always better to use more complex feature sets?

As an example, consider fitting OLS with polynomial features up to a specified degree with 21 data points. If we set degree to be 20, we can fit all data points perfectly and obtain zero residuals (shown in Figure **??**) - but the fitted curve does not pass the sanity check, and will almost certainly perform poorly on new data.

This example illustrates the following two questions we often ask in machine learning.

1. How well will our decision rule perform on new data?

2. How can we choose among a set of possible decision rules that we have already fit to training data? For example, in the polynomial OLS example, how do we choose the max degree of the polynomial features? In ridge regression, how do we choose the hyperparameter λ?

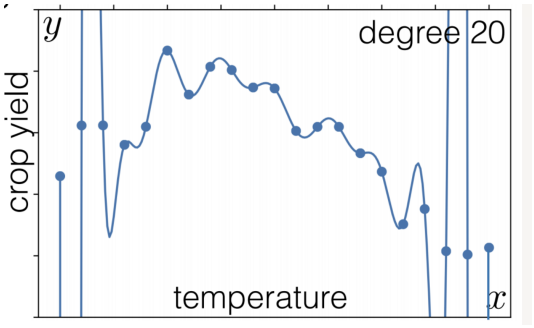


Figure 1: Degree-20 polynomial model.

As assumed throughout this course, our definition of "doing well" is to minimize the expected risk of a new data point (X, Y), i.e., minimizing $\mathbb{E}(L(Y, h_{\mathcal{D}}(X)))$. If we are able to compute this value for arbitrary $h_{\mathcal{D}}$, we can easily answer the above two questions - the performance is the expected loss, and we choose the decision rule that minimizes the expected loss. Of course, the problem is we do not know the distribution of (X, Y) and have to estimate the risk.

To this end, we start by recalling the following tools we have.

Law of Large Numbers. Given i.i.d. random variables $Z^{(1)}, Z^{(2)}, \cdots, Z^{(n)}$ with $\mathbb{E}(|Z^{(1)}|) < \infty$, then with probability 1 we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} Z^{(n)} = \mathbb{E}(Z^{(1)}).$$

We will also use the following assumption: Given $(X^{(n)}, Y^{(n)})$ i.i.d. across $1 \leq n \leq N+1$, for f with $\mathbb{E}(|f(X^{(1)}, Y^{(1)})|) < \infty$, we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(X^{(n)}, Y^{(n)}) = \mathbb{E}(f(X^{(N+1)}, Y^{(N+1)}))$$

2 Empirical Risk and Validation Data

As a starting point, what if we estimate the risk with the empirical average loss over the training data: $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$? To illustrate the problem with this idea, recall the degree-20 polynomial model in our previous example: in the training dataset, the mean squared loss is 0 as the model fits the data perfectly, but we certainly do not expect the expected risk is zero for future data points. Note that the problem is not data size: even if we have a very large dataset, we can define a decision rule that simply memorizes the training set, but it certainly won't generalize well.

So what went wrong here? Actually, we cannot use Law of Large Number because $L(Y^{(i)}, h_{\mathcal{D}}(X^{(i)}))$ are **not i.i.d.** - because $h_{\mathcal{D}}$ is a function of the entire training set.

To fix this problem, we will use *validation data*. Consider M new data points

$$\mathcal{D}' = \{(X^{(m)}, Y^{(m)})\}_{m=N+1}^{N+M}$$

and we estimate the risk with the empirical average loss over the validation data

$$\frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)}))$$

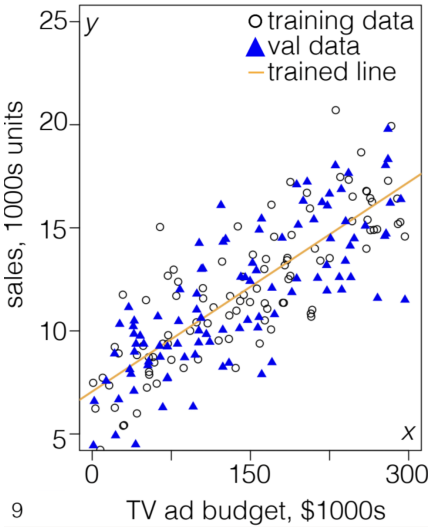
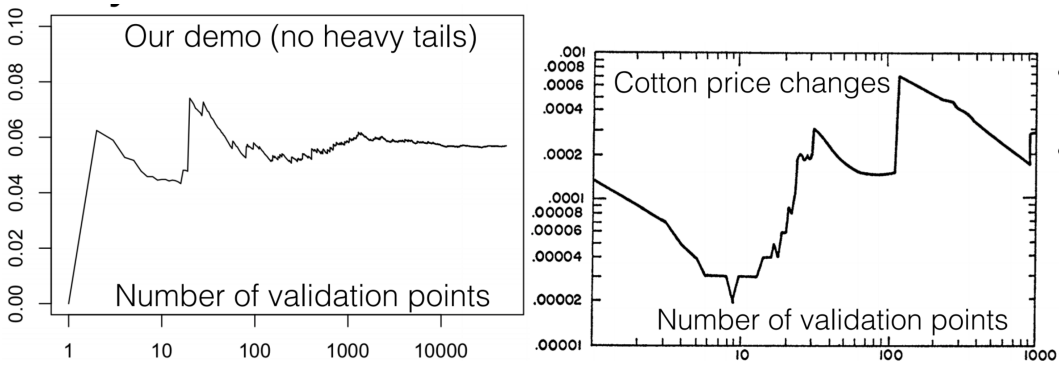


Figure 2: Example of training and validation set for a regression model.

In practice, the validation data can be collected separately from the training data, or we can randomly partition the available data to the training and validation set. But in both cases, we need the i.i.d. assumption.

To illustrate the importance of the i.i.d. condition, consider the following example. We are interested in predicting people's salary from their email (so our feature is the email, and our dependent variable is the sender's salary). Unfortunately, we only have 100 volunteers. So we use 1000 emails for each volunteer to obtain $100,000$ data points total, split into training and validation data, train our model on the training data, and report the empirical loss on the validation data. The problem with this approach is our validation data is not i.i.d. with the training data: any data point in the validation set likely shares the same y (salary) with some data point in the training set.

Another point we want to emphasize is that the x we want to perform inference on must also come from the same distribution as our validation set. Consider the model shown in Figure **??**. The model seems to work well in validation data. Now, suppose we want to know the TV ad budgets in the $\$300,000$ to $\$400,000$ range. The problem is the x values we want to make predictions on are not from the same distribution as the validation x, as the x in validation are drawn from $\$0$ to $\$300,000$. In practice, it is often difficult to predict or evaluate outside the observed data without making any other assumptions.



(a) Normal distribution (b) Heavy-tailed distribution

Figure 3: Standard deviation estimates for different distributions.

3 Size of Validation Data

Now we described how to use validation data to evaluate our model, the next question we ask is how we know we have enough validation data - in other words, how can we quantify the uncertainty in our estimate of the empirical risk?

Recall that for i.i.d. random variables $Z^{(1)}, Z^{(2)}, \cdots, Z^{(N)}$ wih finite mean and variance, the variance of their empirical average is $\mathrm{Var}(Z^{(1)})/N$, and its squared root is called the *standard error*. Of course, in our use case, we do not know the variance of loss of a new data point, so we will need to estimate it from our dataset as well.

One problem that might arise is *heavy-tailed data*. Our calculation and the Law of Large Numbers assume the expectation and variance of loss exist. However, for heavy-tailed distributions, mass in tails does not decay exponentially fast, and the mean or variance of the distribution might not exist. Some examples of heavy-tailed distributions include

1. Daily price changes of certain commodity,
2. Exchange rates,
3. Quiet periods between transmissions for a networked computer terminal,

What happen in those cases?

Figure **??** shows the estimated standard deviation as a function of the number of samples when both mean and variance exist. We see that as we include more samples, the estimated standard deviation converges to the true standard deviation. Figure **??** shows the estimated standard deviation when the variance does not exist.

We can see the big jumps in the graph: this is because once in a while, we will sample a data point with large magnitude that causes the estimated standard deviation to increase significantly; of course, since we know the variance of the distribution does not exist, this estimation procedure would never be able to converge. The takeaway here is that we need to check our assumptions before running any evaluation metrics: without checking for heavy tails and whether the variance exists, simply reporting the average empirical loss over validation data might not be meaningful since we cannot quantify uncertainty with our estimate.