EdX and its Members use cookies and other tracking technologies for performance, analytics, and marketing purposes. By using this website, you accept this use. Learn more about these technologies in the Privacy Policy.

4. Empirical Risk
the Objective: Empirical Risk

7:06 / 7:06 Speed 1.50x

and we are trying to see how far we
can go with linear regression.
So for now what we will assume, that this is--
this empirical risk-- is what we want to optimize.
That we are given our training examples, like the one written
above, feature vectors and the values and labels,
and we are trying to find the status which would result
in the smallest empirical risk.

End of transcript. Skip to the start.

Video
Download video file

Transcripts
Download SubRip (.srt) file

Download Text (.txt) file

Compute Hinge Loss

1/1 point (graded)

The empirical risk R_n is defined as

$$R_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \text{Loss}\left(y^{(t)} - \theta \cdot x^{(t)}\right)$$

where $\left(x^{(t)}, y^{(t)}\right)$ is the tth training example (and there are n in total), and Loss is some loss function, such as hinge loss.

Recall from a previous lecture that the definition of hinge loss:

$$\text{Loss}_h(z) = \begin{cases} 0 & \text{if } z \geq 1 \\ 1-z, & \text{otherwise} \end{cases}.$$

In this problem, we calculate the empirical risk with hinge loss when given specific θ and $\left\{\left(x^{(t)}, y^{(t)}\right)\right\}_{t=1,\ldots,n}$. Assume we have 4 training examples (i.e. $n = 4$), where $x^{(t)} \in \mathbb{R}^3$ and $y^{(t)}$ is a scalar. The training examples $\left\{\left(x^{(t)}, y^{(t)}\right)\right\}_{t=1,2,3,4}$ are given as follow:

	$x^{(t)}$	$y^{(t)}$
$\left(x^{(1)}, y^{(1)}\right)$	$[1,0,1]^T$	2
$\left(x^{(2)}, y^{(2)}\right)$	$[1,1,1]^T$	2.7
$\left(x^{(3)}, y^{(3)}\right)$	$[1,1,-1]^T$	-0.7
$\left(x^{(4)}, y^{(4)}\right)$	$[-1,1,1]^T$	2

Also, we have $\theta = [0,1,2]^T$.

Compute the value of

$$R_n(\theta) = \frac{1}{4}\sum_{t=1}^{4} \text{Loss}_h\left(y^{(t)} - \theta \cdot x^{(t)}\right).$$

1.25 ✔ **Answer:** 1.25

Solution:

From the given table, we can calculate that

$$\begin{aligned} y^{(1)} - \theta \cdot x^{(1)} &= 0, \\ y^{(2)} - \theta \cdot x^{(2)} &= -0.3, \\ y^{(3)} - \theta \cdot x^{(3)} &= 0.3, \\ y^{(4)} - \theta \cdot x^{(4)} &= -1. \end{aligned}$$

Thus we have

$$R_n(\theta) = \frac{1}{4}\sum_{t=1}^{4} \text{Loss}_h\left(y^{(t)} - \theta \cdot x^{(t)}\right) = 1.25$$

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Compute Squared Error Loss

1/1 point (graded)

Now, we will calculate the empirical risk with the squared error loss. Remember that the squared error loss is given by $\text{Loss}(z) = \frac{z^2}{2}$.

The 4 training examples are as in the previous problem:

$x^{(t)}$	$y^{(t)}$

$\left(x^{(1)}, y^{(1)}\right)$	$[1,0,1]^T$	2
$\left(x^{(2)}, y^{(2)}\right)$	$[1,1,1]^T$	2.7
$\left(x^{(3)}, y^{(3)}\right)$	$[1,1,-1]^T$	-0.7
$\left(x^{(4)}, y^{(4)}\right)$	$[-1,1,1]^T$	2

As in the problem above, we have $\theta = [0,1,2]^T$.

Compute the value of

$$R_n(\theta) = \frac{1}{4}\sum_{t=1}^{4} \text{Loss}\left(y^{(t)} - \theta \cdot x^{(t)}\right)$$

.

0.1475 ✔ **Answer:** 0.1475

Solution:

$$R_n(\theta) = \frac{1}{4}(0 + 0.045 + 0.045 + 0.5) = 0.1475.$$

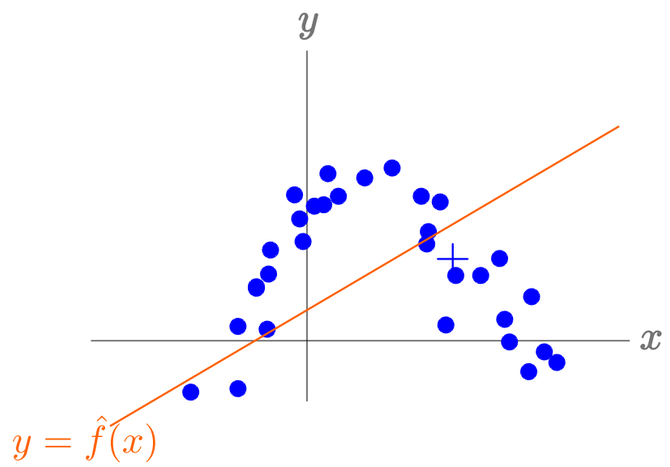
Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Geometrically Identifying Error

1/1 point (graded)

What type of error does the figure below depict? The blue dots are the training examples, and the red line is the predictor $\hat{f}(x)$



- [x] Structural error ✔
- [ ] Estimation error

✔

Solution:

Here, structural error occurs because the true underlying relationship is non-linear but the regression function is linear.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Increasing the Number of Training Examples

1/1 point (graded)
If we increase n, the number of training examples, which of the following types of errors decreases?

- [ ] Structural error
- [x] Estimation error ✔

✔

Solution:

The larger the training set is, the smaller the estimation error will be. Structural error occurs when the true underlying relationship is highly non-linear, so it is not relevant to increasing n.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Empirical Risk and Model Performance

1/1 point (graded)
If we are given a large amount of training data and successfully obtain 0 empirical risk, is the model we learned guaranteed to perform well on the test data? (Assume training and test data are drawn from the same distribution.)

- Yes.
- No. ✔

Solution:

No. Obtaining 0 empirical risk for a large amount of data means that it is possible that the model is overfitted.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

(Optional) Error decomposition and the bias-variance trade-off

(Optional) Error decomposition and the bias-variance trade-off

In the lecture, we intuitively defined the structural and estimation mistakes. Here we provide a more formal definition of these ideas for a regression problem.

Suppose we want to learn the relationship between random variables $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, where the ground truth relationship is $y = f(x)$. Of course f is unkown to us, but we observe a training set (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) and we hope to find a function $\hat{f}$ to approximate the true funtion f.

However, our observed data might not be 100% accurate as there can be many kinds of noise and uncertainty containing in the data. Thus, we further assume a random noise variable ϵ is added on top of y:

$$y = f(x) + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$

We learned from the lecture that we can find $\hat{f}$ by minimizing the empirical risk on the training set. As we know the training set is a random observation of the underlying relationship and contains noise, different training set will give us different estimator $\hat{f}(x)$. And hence we can define $\mathbb{E}[\hat{f}(x)]$ to be the expected estimator over all possible training sets.

Now let's look at when we have a new x with unkonwn y, what is the expected prediction error looks like for our estimator given all possible training sets:

$$\begin{aligned}\mathbb{E}[(y - \hat{f}(x))^2] &= \mathbb{E}[(f(x) + \epsilon - \hat{f}(x))^2] \\ &= (f(x) - \mathbb{E}[\hat{f}(x)])^2 + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + \mathbb{E}[\epsilon^2]\end{aligned}$$

We skip some derivations of this result, can you get this result on your own?

As we can see, there are three terms in this error decomposition:

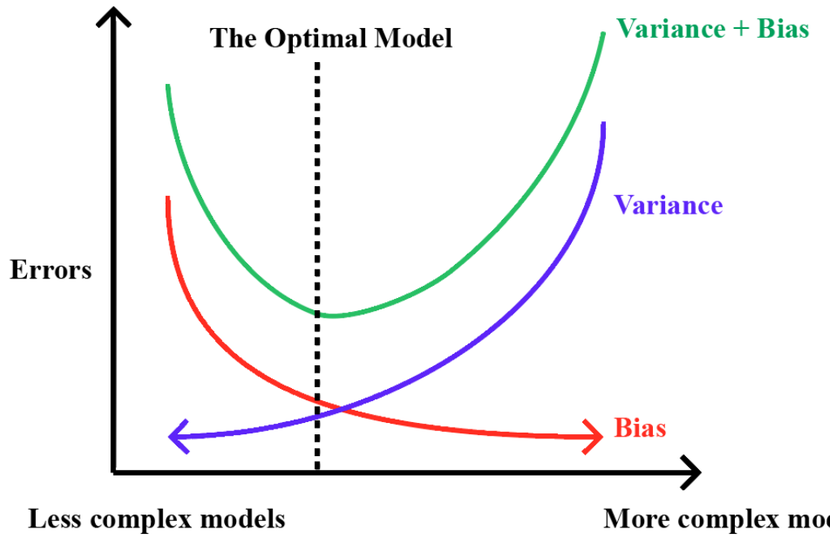
The first term is the square of the difference between the true $f(x)$ and the expected estimation over all possible training sets. This term is usually called bias and it describes how much the average estimator fitted over all datasets deviates from the underlying true $f(x)$. This corresponds to the structural mistake in the lecture.

The second term is the variacne of the estimator (recall variance definition in statistics). It describes on average how much a single estimator deviates from the expected estimator over all data sets. This corresponds to the estimation error in the lecture.

The thrid term $\mathbb{E}[\epsilon^2] = \sigma^2$ is the error from the inherent noise of the data and we can do nothing to minimize it, thus it is sometimes called irreducible error. The irreducible error gives a lower bound on the expected prediction error.

Question for you to think: Apparently, the empricial risk is just an approximation of the true risk (i.e. $\mathbb{E}\left[\left(f\left(x\right)-\hat{f}\left(x\right)\right)^{2}\right]$). If we were able to learn the model by minimizing the true risk, which part of the error decomposition will become 0?

The task of supervised learning is to reduce the bias and variance at the same time, but because of the noise in the training data, it is not possible to simultaneously minimize these two sources of errors. This is known as the bias-variance trade-off.



To reduce bias, we can assume a more complex hypothesis space and fit a more powerful model. The model will be able to fit even the noise in the training set. However, this increases the error from variance becuase given another training set, the randomness of the noise will results in a very different model. This situation is often called 'overfitting'. On the other hand, we can have a more simple model to reduce the variance, but this can make the bias very large. For example, in the extreme case, let our model be $\hat{f}\left(x\right)=c$, where c is a constant. This will give us 0 variance but you can imagine it can hardly make any correct predictions. This situation is known as 'underfitting'.

In a few pages, you will see how regularization can be used to restrict the complexity for linear regression so that we will be able to search for a sweet spot where the total error from variance and bias is the minimum.

Hide

Discussion

Hide Discussion

Topic: Unit 2 Nonlinear Classification, Linear regression, Collaborative Filtering (2 weeks):Lecture 5. Linear Regression / 4. Empirical Risk

Add a Post

All Posts

Empirical Risk and Model Performance

discussion posted a day ago by **am2384**

This question is poorly written, may even be erroneous.

(1) It explicitly mentions "**large**" training data. The chances of overfitting decrease rapidly as the size of the training data increases. It is very difficult to have overfitting for a **large** training data; hence, if we have negligible loss on a "**large**" training data; it is a very good guarantee that the model will generalize very well to the test data as well.

(2) The question only asks for a "guarantee for performing well on the test data" without quantitatively specifying how well. Another poor wording/ambiguity.

The question makes much more sense if "large" is replaced by "small".

This post is visible to everyone.

Add a Response

2 responses

ptressel (Community TA)
about 14 hours ago

I agree -- I'm boycotting this question (for now) over the feeling that it's a trick question. 😉 Does "guaranteed" mean absolutely 100%? 🤔 "Large" compared to what? Usually that would be compared to the model complexity. But it might mean "large" in some informal sense, like "a petabyte is large". I will yield and hit the submit button sooner or later... 😝

Least square estimator is a consistent estimator. Remove the question.

posted about 5 hours ago by **maximoff**

Add a comment

davidstuartbruce
about a minute ago

I agree that it is not worded very well, but the point it seems to be making is valid. A larger data set reduces overfitting, but it is always possible to devise such a complex model that it is fitting both the "true" $f(x)$ and the error. Fitting to the error is not going to help performance on the test set. Also, I'm always wary about answering "true" to anything involving a guarantee, though such skepticism is not infallible.

David

Add a comment

Showing all responses

Add a response:

Preview

Submit

Learn About Verified Certificates

© All Rights Reserved