NEURAL NETWORKS AND DEEP LEARNING CST 395 CS 5TH SEMESTER HONORS COURSE- Dr Binu V P, 9847390760

CS 5th Semester Honors course for the Computer Science at KTU- Dr Binu V P

Estimators, Bias , Variance and Consistency

October 01, 2022

The field of statistics gives us many tools that can be used to achieve the machine learning goal of solving a task not only on the training set but also to generalize.Foundational concepts such as parameter estimation, bias and variance are useful to formally characterize notions of generalization, underfitting and overfitting.

Point Estimation

Point estimation is the attempt to provide the single “best” prediction of some quantity of interest. In general the quantity of interest can be a single parameter or a vector of parameters in some parametric model, such as the weights in our linear regression example , but it can also be a whole function.

In order to distinguish estimates of parameters from their true value, our convention will be to denote a point estimate of a parameter θ by $\hat{\theta}$.

Let $\{x^{(1)}, \ldots, x^{(m)}\}$ be a set of m independent and identically distributed (i.i.d.) data points. A point estimator or statistics is any function of the data:

The definition does not require that g return a value that is close to the true θ or even that the range of g is the same as the set of allowable values of θ.This definition of a point estimator is very general and allows the designer of an estimator great flexibility. While almost any function thus qualifies as an estimator, a good estimator is a function whose output is close to the true underlying θ that generated the training data.

For now, we take the frequentist perspective on statistics. That is, we assume that the true parameter value θ is fixed but unknown, while the point estimate $\hat{\theta}$ is a function of the data. Since the data is drawn from a random process, any function of the data is random. Therefore $\hat{\theta}$ is a random variable.Point estimation can also refer to the estimation of the relationship between input and target variables. We refer to these types of point estimates as **function estimators**.

Function Estimation

As we mentioned above, sometimes we are interested in performing function estimation (or function approximation). Here we are trying to predict a variable y given an input vector x. We assume that there is a function $f(x)$ that describes the approximate relationship between y and x. For example, we may assume that $y = f(x) + \epsilon$, where ϵ stands for the part of y that is not predictable from x.

In function estimation, we are interested in approximating f with a model or estimate $\hat{f}$. Function estimation is really just the same as estimating a parameter θ; the function estimator $\hat{f}$ is simply a point estimator in function space. The linear regression example and the polynomial regression example are both examples of scenarios that may be interpreted either as estimating a parameter w or estimating a function $\hat{f}$ mapping from x to y.

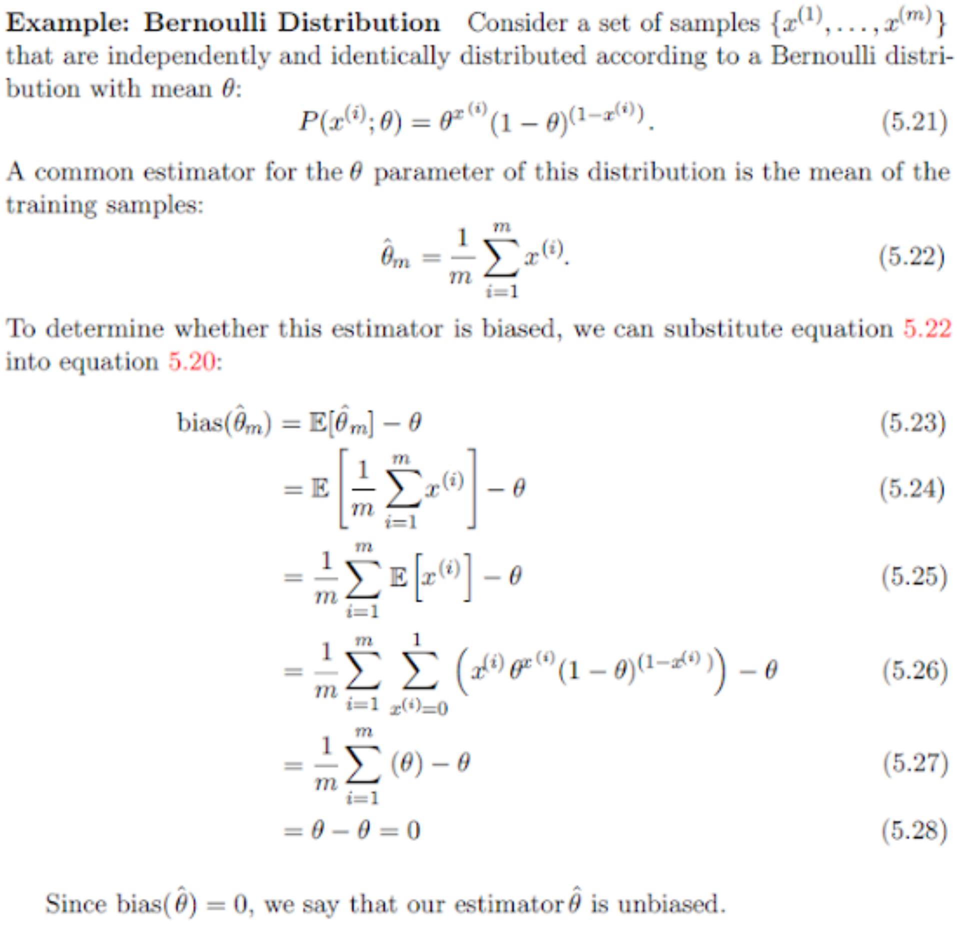
We now review the most commonly studied properties of point estimators and discuss what they tell us about these estimators.

Bias

The bias of an estimator is defined as:

$bias(\hat{\theta}_m) = E(\hat{\theta}_m) - \theta$

where the expectation is over the data (seen as samples from a random variable) and θ is the true underlying value of θ used to define the data generating distribution.An estimator $\hat{\theta}_m$ is said to be unbiased if $bias(\hat{\theta}_m = 0$, which implies that $E(\hat{\theta}_m) = \theta$. An estimator $\hat{\theta}_m$ is said to be asymptotically unbiased if $lim_{m \to \infty} bias(\hat{\theta}_m) = 0$, which implies that $lim_{m \to \infty} E(\hat{\theta}_m) = \theta$.



Example: Bernoulli Distribution Consider a set of samples $\{x^{(1)}, \dots, x^{(m)}\}$ that are independently and identically distributed according to a Bernoulli distribution with mean θ:

$$P(x^{(i)}; \theta) = \theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}. \tag{5.21}$$

A common estimator for the θ parameter of this distribution is the mean of the training samples:

$$\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}. \tag{5.22}$$

To determine whether this estimator is biased, we can substitute equation 5.22 into equation 5.20:

$$\begin{aligned}
\text{bias}(\hat{\theta}_m) &= \mathbb{E}[\hat{\theta}_m] - \theta && (5.23)\\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \theta && (5.24)\\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right] - \theta && (5.25)\\
&= \frac{1}{m}\sum_{i=1}^{m} \sum_{x^{(i)}=0}^{1} \left(x^{(i)}\theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}\right) - \theta && (5.26)\\
&= \frac{1}{m}\sum_{i=1}^{m} (\theta) - \theta && (5.27)\\
&= \theta - \theta = 0 && (5.28)
\end{aligned}$$

Since bias$(\hat{\theta}) = 0$, we say that our estimator $\hat{\theta}$ is unbiased.

Variance and Standard Error

Another property of the estimator that we might want to consider is how much we expect it to vary as a function of the data sample. Just as we computed the expectation of the estimator to determine its bias, we can compute its variance.The variance of an estimator is simply the variance

$$Var(\hat{\theta})$$

where the random variable is the training set. Alternately, the square root of the variance is called the , denoted **standard error** $SE(\hat{\theta})$.

The variance or the standard error of an estimator provides a measure of how we would expect the estimate we compute from data to vary as we independently resample the dataset from the underlying data generating process. Just as we might like an **estimator to exhibit low bias we would also like it to have relatively low variance.**

When we compute any statistic using a finite number of samples, our estimate of the true underlying parameter is uncertain, in the sense that we could have obtained other samples from the same distribution and their statistics would have been different. The expected degree of variation in any estimator is a source of error that we want to quantify.The standard error of the mean is given by

$$SE(\hat{\mu}_m) = \sqrt{Var\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}}$$

where σ^2 is the true variance of the samples x^i . The standard error is often estimated by using an estimate of σ. Unfortunately, neither the square root of the sample variance nor the square root of the unbiased estimator of the variance provide an unbiased estimate of the standard deviation. Both approaches tend to underestimate the true standard deviation, but are still used in practice. The square root of the unbiased estimator of the variance is less of an underestimate. For large m, the approximation is quite reasonable.

The standard error of the mean is very useful in machine learning experiments. We often estimate the generalization error by computing the sample mean of the error on the test set. The number of examples in the test set determines the accuracy of this estimate. Taking advantage of the central limit theorem, which tells us that the mean will be approximately distributed with a normal distribution, we can use the standard error to compute the probability that the true expectation falls in any chosen interval. For example, the 95% confidence interval centered on the mean $\hat{\mu}_m$ is

$$(\hat{\mu}_m - 1.96SE(\hat{\mu}_m), \hat{\mu}_m + 1.96SE(\hat{\mu}_m))$$

under the normal distribution with mean $\hat{\mu}_m$ and variance $SE(\hat{\mu}_m)^2$. In machine learning experiments, it is common to say that algorithm A is better than algorithm B if the upper bound of the 95% confidence interval for the error of algorithm A is less than the lower bound of the 95% confidence interval for the error of algorithm B.

Trading off Bias and Variance to Minimize Mean Squared Error

Bias and variance measure two different sources of error in an estimator. Bias measures the expected deviation from the true value of the function or parameter.Variance on the other hand, provides a measure of the deviation from the expected estimator value that any particular sampling of the data is likely to cause.

What happens when we are given a choice between two estimators, one with more bias and one with more variance? How do we choose between them? The most common way to negotiate this trade-off is to use cross-validation.Empirically, cross-validation is highly successful on many real-world tasks.

Alternatively, we can also compare the mean squared error (MSE) of the estimates:

$MSE = E[(\hat{\theta}_m - \theta)^2] = Bias^2(\hat{\theta}_m) + Var(\hat{\theta}_m)$

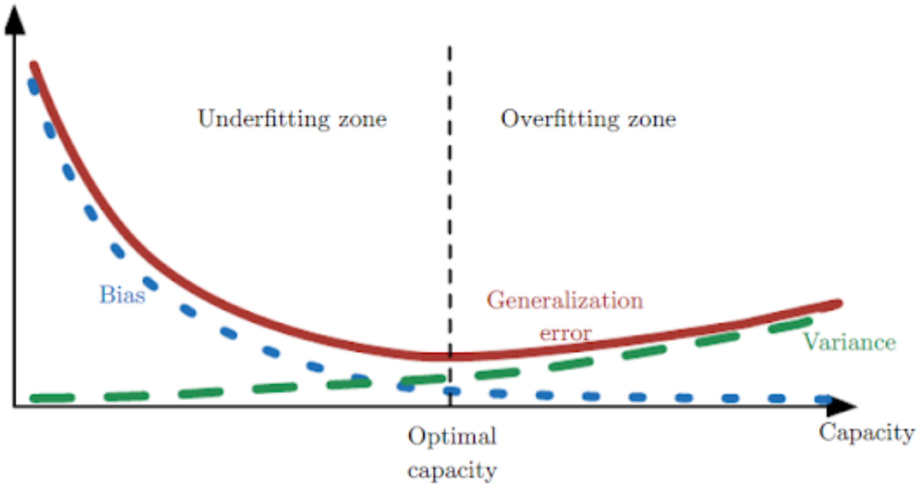
Proof:

$E[(\hat{\theta}_m - \theta)^2] = E[(\hat{\theta}_m)^2] + \theta^2 - 2E(\hat{\theta}_m)\theta$

$Bias^2(\hat{\theta}_m) = (E(\hat{\theta}_m) - \theta)^2 = E^2(\hat{\theta}_m) + \theta^2 - 2E(\hat{\theta}_m)\theta$

$Var(\hat{\theta}_m) = E[(\hat{\theta}_m)^2] - E^2[(\hat{\theta}_m)]$
it is noted that $Bias^2(\hat{\theta}_m) + Var(\hat{\theta}_m) = MSE$

The MSE measures the overall expected deviation—in a squared error sense—between the estimator and the true value of the parameter θ. As is clear from above equation , evaluating the MSE incorporates both the bias and the variance.Desirable estimators are those with small MSE and these are estimators that manage to keep both their bias and variance somewhat in check.



As capacity increases (x-axis), bias (dotted) tends to decrease and variance (dashed) tends to increase, yielding another U-shaped curve for generalization error (bold curve). If we vary capacity along one axis, there is an optimal capacity, with underfitting when the capacity is below this optimum and overfitting when it is above. This relationship is similar to the relationship between capacity, underfitting, and overfitting,

The relationship between bias and variance is tightly linked to the machine learning concepts of capacity, underfitting and overfitting. In the case where generalization error is measured by the MSE (where bias and variance are meaningful components of generalization error), increasing capacity tends to increase variance and decrease bias. This is illustrated in above figure , where we see again the U-shaped curve of generalization error as a function of capacity.

Consistency

So far we have discussed the properties of various estimators for a training set of fixed size. Usually, we are also concerned with the behavior of an estimator as the amount of training data grows. In particular, we usually wish that, as the number of data points m in our dataset increases, our point estimates converge to the true value of the corresponding parameters. More formally, we would like that

$plim_{m \to \infty} = \hat{\theta}_m = \theta$

The symbol $plim$ indicates convergence in probability, meaning that for any $\epsilon > 0$,$P(|\hat{\theta}_m - \theta| > \epsilon) \rightarrow 0$ as $m \rightarrow \infty$. The condition described by equation is known as **consistency**. It is sometimes referred to as weak consistency, with strong consistency referring to the almost sure convergence of $\hat{\theta}$ to θ. Almost sure convergence of a sequence of random variables $x^{(1)}, x^{(2)}, \ldots$. to a value x occurs when $p(lim_{m \rightarrow \infty} x(m) = x) = 1$.

Consistency ensures that the bias induced by the estimator diminishes as the number of data examples grows. However, the reverse is not true—asymptotic unbiasedness does not imply consistency. For example, consider estimating the mean parameter μ of a normal distribution $N(x; \mu, \sigma^2)$, with a dataset consisting of m samples: $\{x^{(1)}, \ldots, x^{(m)}\}$. We could use the first sample $x^{(1)}$ of the dataset as an unbiased estimator:$\hat{\theta} = x^{(1)}$. In that case, $E(\hat{\theta}_m) = \theta$ so the estimator is unbiased no matter how many data points are seen. This, of course, implies that the estimate is asymptotically unbiased. However, this is not a consistent estimator as it is not the case that $\hat{\theta}_m \rightarrow \theta$ as $m \rightarrow \infty$.

Theme images by Michael Elkan

Popular posts from this blog

NEURAL NETWORKS AND DEEP LEARNING CST 395 CS 5TH SEMESTER HONORS COURSE NOTES - Dr Binu V P, 9847390760

October 03, 2022

About Me Syllabus Question Paper Dec 2022 Module 1 (Basics of Machine Learning) Overview of Machine Learning Machine Learning Algorithm Linear Regression Capacity, Overfitting and Underfitting Regularization Hyperparameters and Validatio …

READ MORE

Machine Learning Algorithm

October 01, 2022

A machine learning algorithm is an algorithm that is able to learn from data. But what do we mean by learning? Mitchell (1997) provides the definition “A computer program is said to learn from experience E with respect to some class of tasks T and …

READ MORE

Syllabus

October 01, 2022

Syllabus Module - 1 (Basics of Machine Learning) Machine Learning basics - Learning algorithms - Supervised, Unsupervised, Reinforcement, overfitting, Underfitting, Hyper parameters and Validation sets, Estimators -Bias and Variance. Challenges …

READ MORE