# Statistics for Deep Learning

Sung-Yub Kim

Dept of IE, Seoul National University

January 14, 2017

- **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.
To distinguish estimates of parameters from their true value, our convention will be dentoe a point estimate of a parameter $\theta$ by $\hat{\theta}$
A good estimator is a function whose output is **close to the true underlying** $\theta$ that generated the training data.

- **Funtion Estimation**

Sometimes we are interested in performing function approximation. Here we trying to predict a variable **y** given **x**. Therefore, we may assume that

$$\mathbf{y} = f(\mathbf{x}) + \epsilon \tag{1}$$

- **Bias**

The bias of an estimator is defined as

$$bias(\hat{\theta_m}) = \mathbb{E}[\hat{\theta_m}] - \theta \tag{2}$$

- **Unbiased**

An estimator $\hat{\theta_m}$ is said to be **unbiased** if

$$bias(\hat{\theta_m}) = 0 \tag{3}$$

,which implies that $\mathbb{E}[\hat{\theta_m}] = \theta$.

- **Asymptotically Unbiased**

An estimator $\hat{\theta_m}$ is said to be **asymptotically unbiased** if

$$\lim_{m \to \infty} bias(\hat{\theta_m}) = 0 \tag{4}$$

, which implies that $\lim_{m \to \infty} \mathbb{E}[\hat{\theta_m}] = \theta$

- **Variance**

The variance of an estimator is simply the variance

$$Var(\hat{\theta}) \tag{5}$$

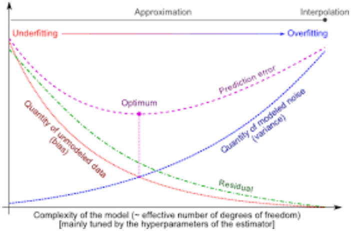
where the random variable is the training set. Alternately, the square root of the variance is called the **standard error**, denoted $SE(\hat{\theta})$.
The variance of an estimator provides a measure of how we would expect the estimate we compute from data to **vary as we independently resample the dataset** from underlying data-generating process.

- **Bias-Variance Tradeoff**

The realtion between bias and variance can be showed as term of **mean squared error**(MSE) of the estimtes:

$$MSE = \mathbb{E}[(\hat{\theta}_m - \theta)] = Bias(\hat{\theta}_m)^2 + Var(\hat{\theta}_m) \tag{6}$$

Desirable estimators are those with small MSE and these are estimators that manage to keep both their bias and variance somewhat in check. Usually, we measure generalization error is measured by the MSE.



- **Weak Consistency**

Usually, we are also concerned with the behavior of an estimator as **the amount of training data grows**. We wish that, as the number of data points $m$ in our dataset increases, our point converge to the true value of the corresponding parameters. More formally, we define an estimate of parameter is weak consistent if

$$p\lim_{m\to\infty} \hat{\theta_m} = \theta \tag{7}$$

The symbol $p\lim$ indicates **convergence in probability**, meaning that for any $\epsilon > 0$, $\mathbb{P}[|\hat{\theta_m} - \theta| > \epsilon] \to 0$ as $m \to 0$.

- **Strong Consistency**

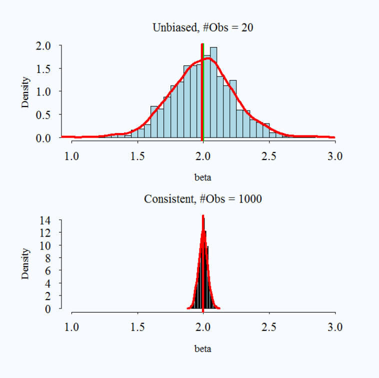
Also, we can define strong consistency like

$$\mathbb{P}[\lim_{m\to\infty} \hat{\theta_m} = \theta] = 1 \tag{8}$$

This equation also means that **almost surely convergence**,
$\mathbb{P}[\liminf_{m\to\infty}\{\omega \in \Omega : |\theta_m(\omega) - \hat{\theta}(\omega)| < \epsilon\}] = 1, \forall\epsilon > 0$

- Consistency and Asymptotically Unbiased

Consistency ensures that bias induced by the estimator diminishes as the number of data example grows. But, the reverse in not true, since even if we have enough examples, we can make proabability bigger than a positive number by controlling $\epsilon$.



- **Maximum Likelihood Estimation**

Let $p_{model}(\mathbf{x};\theta)$ be a parametric family of probability distributions over the same space indexed by $\theta$. The Maximum Likelehood Estimator(MLE) for $\theta$ is defined as

$$\theta_{MLE} = arg\max_{\theta} p_{model}(\mathbb{X};\theta) = arg\max_{\theta}\prod_{i=1}^{m} p_{model}(\mathbf{x}^{(i)};\theta) \tag{9}$$

and this maximization problem is equivalent to

$$arg\max_{\theta}\sum_{i=1}^{m}\log p_{model}(\mathbf{x}^{(i)};\theta) = arg\max_{\theta}\mathbb{E}_{\mathbf{x}\sim\hat{p}_{data}}[\log p_{model}(\mathbf{x};\theta)] \tag{10}$$

where $\hat{p}_{data}$ is a empirical distribution. And this can be interpreted as

$$arg\min_{\theta} D_{KL}(\hat{p}_{data}\|p_{model}) = arg\max_{\theta}\mathbb{E}_{\mathbf{x}\sim\hat{p}_{data}}[\log\hat{p}_{data}(x)-\log p_{model}(\mathbf{x};\theta)] \tag{11}$$

since this minimization problem is equivalent to

$$-arg\min_{\theta}\mathbb{E}_{\mathbf{x}\sim\hat{p}_{data}}[\log p_{model}(\mathbf{x})] \tag{12}$$

- **Minimizing the Cross Entropy is all we do**

Any loss consisting of a **negative log-likelihood** is a cross-entropy between the empirical distribution defined by the training set and the probability distribution defined by model. For example, MSE is the cross-entropy between the empirical distribution and a Gaussian model.

- MLE is a consistent estimator.

If we assume that the true distribution $p_{data}$ must lie within the model family $p_{model}$ and the true distribution $p_{data}$ must correspond to exactly one value of $\theta$.

- Consistent estimator has statistical efficiency.

It means that one consistent estimator may obtain lower generalization error for a fixed number of samples or requires fewer examples to obtain a fixed level of generalization error.

- **Cramer-Rao Bound**

Cramer-Rao lower bound show that no consistent estimator has a lower MSE than the MLE.

- Differences between Frequentists and Bayesians

In frequentist statistics, we have no asssumptions on $\theta$ except $\theta$ is fixed. But in bayesian statistics, we assume that $\theta$ is not fixed and have a distribution. We call this distribution **prior**. In ML, we set this prior high entropy to reflect a high degree of uncertainty in the value of $\theta$ before observing any data.

- Differences in estimation

Since estimation is a process of minimization of cross-entropy, in MLE we need to consider $\theta$ in finite vector space. But bayesian assumes that $\theta$ has a distribution we need to consider **distribution of** $\theta$ in infinite dimensional vector space(function space).

Since we can parameterize prior and posterior distribution, we can make a single-point estimation. And this is so called Maximum A Posteriori(MAP) estimation.

$$\theta_{MAP} = arg \max_{\theta} p(\theta|\mathbf{x}) = arg \max_{\theta} \log p(\mathbf{x}|\theta) + \log p(\theta) \tag{13}$$

and it can be interpreted as minimzation of sum of **standard log-likelihood** and **log-prior**. And this is correspond to weight-decay.