

CSE 6740 Lecture 3

How Do I Learn a Simple Model? (Probability and inference)

Alexander Gray

agray@cc.gatech.edu

Georgia Institute of Technology

Today

1. Asymptotics and point estimation (*What is “estimation/learning”?*), cont'd.
2. Confidence intervals (*How good is the estimation/learning?*)

Asymptotic theory and point estimation

What is “estimation/learning”? What happens as you get more data? Why is the sample mean a good estimator?

Markov's Inequality

Theorem (*Markov's inequality*): Suppose X is a non-negative random variable and $\mathbb{E}(X)$ exists. Then for any $t > 0$,

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}. \quad (1)$$

Markov's Inequality: Proof

Since $X > 0$,

$$\mathbb{E}(X) = \int_0^{\infty} x f(x) dx \quad (2)$$

$$= \int_0^t x f(x) dx + \int_t^{\infty} x f(x) dx \quad (3)$$

$$\geq \int_t^{\infty} x f(x) dx \quad (4)$$

$$\geq t \int_t^{\infty} f(x) dx \quad (5)$$

$$= t\mathbb{P}(X > t). \quad (6)$$

Chebyshev's Inequality

Theorem (*Chebyshev's inequality*): If $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{V}(X)$, then

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad (7)$$

and

$$\mathbb{P}\left(\left|\frac{X - \mu}{\sigma}\right| \geq u\right) \leq \frac{1}{u^2} \quad (8)$$

(or $\mathbb{P}(|Z| \geq u) \leq \frac{1}{u^2}$ if $Z = (X - \mu)/\sigma$).

For example, $\mathbb{P}(|Z| > 2) \leq 1/4$ and $\mathbb{P}(|Z| > 3) \leq 1/9$.

Chebyshev's Inequality: Proof

Using Markov's inequality,

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^2 \geq t^2) \quad (9)$$

$$\leq \frac{\mathbb{E}(X - \mu)^2}{t^2} \quad (10)$$

$$= \frac{\sigma^2}{t^2}. \quad (11)$$

The second part follows by setting $t = u\sigma$.

Chebyshev's Inequality: Example

Suppose we test a classifier on a set of N new examples. Let $X_i = 1$ if the prediction is wrong and $X_i = 0$ if it is right; then $\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i$ is the observed error rate. Each X_i may be regarded as a Bernoulli with unknown mean p ; we would like to estimate this.

How likely is \bar{X}_N to not be within ϵ of p ?

Chebyshev's Inequality: Example

We have that $\mathbb{V}(\bar{X}_N) = \mathbb{V}(X)/N = p(1 - p)/N$ and

$$\mathbb{P}(|\bar{X}_N - p| > \epsilon) \leq \frac{\mathbb{V}(\bar{X}_N)}{\epsilon^2} \quad (12)$$

$$= \frac{p(1 - p)}{N\epsilon^2} \quad (13)$$

$$\leq \frac{1}{4N\epsilon^2} \quad (14)$$

since $p(1 - p) \leq 1/4$ for all p .

For $\epsilon = .2$ and $N = 100$ the bound is .0625.

Hoeffding's Inequality

Similar to Markov's, but tighter.

Theorem (*Hoeffding's inequality*): Let X_1, \dots, X_N be independent observations such that $\mathbb{E}(X_i) = 0$ and $a_i \leq X_i \leq b_i$. Then for any $t > 0, \epsilon > 0$,

$$\mathbb{P} \left(\sum_{i=1}^N X_i \geq \epsilon \right) \leq e^{-t\epsilon} \prod_{i=1}^N e^{t^2(b_i - a_i)^2 / 8}. \quad (15)$$

Hoeffding's Inequality

We are most often interested in this special case: Let $X_1, \dots, X_N \sim \text{Bernoulli}(p)$. Then for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_N - p| > \epsilon) \leq 2e^{-2N\epsilon^2} \quad (16)$$

where $\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i$.

Hoeffding's Inequality: Example

Recall our previous example, where X_1, \dots, X_N Bernoulli(p). Letting $N = 100$ and $\epsilon = .2$, Chebyshev's inequality yielded

$$\mathbb{P}(|\bar{X}_{100} - p| > .2) \leq .0625. \quad (17)$$

Hoeffding's inequality yields the tighter bound

$$\mathbb{P}(|\bar{X}_{100} - p| > .2) \leq 2e^{-2 \cdot 100 \cdot (.2)^2} = .00067. \quad (18)$$

(Weak) Law of Large Numbers

Theorem (*WLLN*): If X_1, \dots, X_N are IID, and $\mathbb{E}(X_i) = \mu$, then $\bar{X}_N \xrightarrow{p} \mu$.

This says that the sample mean \bar{X}_N approaches the true mean μ as N gets large.

WLLN: Proof

To make the proof simpler (though it's not strictly necessary), assume the variance is finite ($\sigma < \infty$). Then using Chebyshev's inequality,

$$\mathbb{P}(|\bar{X}_N - \mu| > \epsilon) \leq \frac{\mathbb{V}(\bar{X}_N)}{\epsilon^2} \quad (19)$$

$$= \frac{\sigma^2}{N\epsilon^2} \quad (20)$$

which approaches 0 as $N \rightarrow \infty$.

Point estimation

We want to make a single best guess at the value of some true quantity θ^* , or simply θ . We call our guess $\hat{\theta}$ or $\hat{\theta}_N$. θ is a fixed quantity; $\hat{\theta}$ depends on the data so it is a random variable.

If X_1, \dots, X_N are IID data points from some distribution F , a point *estimator* $\hat{\theta}_N$ of some parameter θ is some function $g(X_1, \dots, X_N)$. It can also be called a *learner* or *decision rule*.

Sampling distribution

Since $\hat{\theta}_N$ is itself a random variable, it has a distribution. The distribution of $\hat{\theta}_N$ is called the *sampling distribution*. The standard deviation of $\hat{\theta}_N$ is called the *standard error*

$$\text{se} = \text{se}(\hat{\theta}_N) = \sqrt{\mathbb{V}(\hat{\theta}_N)}. \quad (21)$$

We call the estimated standard error $\widehat{\text{se}}$.

Properties of Estimators

What makes for a good estimator?

The *bias* of an estimator is

$$\text{bias}(\hat{\theta}_N) = \mathbb{E}_\theta(\hat{\theta}_N) - \theta. \quad (22)$$

The estimator is *unbiased* if $\mathbb{E}_\theta(\hat{\theta}_N) = \theta$.

A point estimator is *consistent* if $\hat{\theta}_N \xrightarrow{p} \theta$.

Also, we can talk about the *convergence rate* of an estimator. And thus, *optimal* estimators.

Properties of Estimators

Unbiasedness versus consistency: Bias must go to zero to achieve consistency. But just being unbiased doesn't imply consistency. And consistency may only imply asymptotic unbiasedness.

Asymptotic unbiasedness: if we converge in distribution to a distribution whose mean is the right value.

Confidence bands

How good is the estimation/learning? We'd like to know some upper and lower bounds on our estimates.

Confidence Sets

A $1 - \alpha$ *confidence interval* for a parameter θ is an interval $C_N = (a, b)$ where $a = a(X_1, \dots, X_N)$ and $b = b(X_1, \dots, X_N)$ are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_N) \geq 1 - \alpha \quad (23)$$

for all $\theta \in \Theta$. In other words, (a, b) traps θ with probability $1 - \alpha$. We call $1 - \alpha$ the *coverage* of the confidence interval.

If θ is a vector we have a *confidence set*, which is a ball or ellipse.

Hoeffding Confidence Band

Consider a coin flip example, with Bernoulli parameter p . Hoeffding's inequality gives us a simple way to create a confidence band. Fix $\alpha > 0$ and let

$$\epsilon_N = \sqrt{\frac{1}{2N} \log \left(\frac{2}{\alpha} \right)}. \quad (24)$$

By Hoeffding's inequality,

$$\mathbb{P}(|\bar{X}_N - p| > \epsilon_N) \leq 2e^{-2N\epsilon_N^2} = \alpha. \quad (25)$$

The interval $C = (\bar{X}_N - \epsilon_N, \bar{X}_N + \epsilon_N)$ traps the true parameter p with probability $1 - \alpha$.

Asymptotic Normality

An estimator is *asymptotically normal* if

$$\frac{\hat{\theta}_N - \theta}{\text{se}} \rightsquigarrow \mathcal{N}(0, 1). \quad (26)$$

Showing that an estimator is asymptotically normal is one way of obtaining a confidence interval (set) for it.

Normal-based Confidence Interval

Suppose that $\hat{\theta}_N \approx \mathcal{N}(\theta, \widehat{\text{se}}^2)$.

Let $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$, that is, $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ and $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$ where $Z \sim \mathcal{N}(0, 1)$. For a 95% confidence interval, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$, leading to the approximate confidence interval $\hat{\theta}_N \pm 2\widehat{\text{se}}$.

Then $\mathbb{P}_\theta(\theta \in C_N) \rightarrow 1 - \alpha$.

Central Limit Theorem

Theorem (CLT): If X_1, \dots, X_N are IID (with *any* distribution), with mean μ and variance σ^2 , then

$$Z_N = \frac{\bar{X}_N - \mu}{\sqrt{\mathbb{V}(\bar{X}_N)}} = \frac{\sqrt{N}(\bar{X}_N - \mu)}{\sigma} \rightsquigarrow Z \quad (27)$$

where $Z \sim \mathcal{N}(0, 1)$. In other words,

$$\lim_{N \rightarrow \infty} \mathbb{P}(Z_N \leq z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \quad (28)$$

Central Limit Theorem

This says that probability statements about \bar{X}_N can be approximated using a Normal distribution. This is written as

$$Z_N = \frac{\sqrt{N}(\bar{X}_N - \mu)}{\sigma} \approx \mathcal{N}(0, 1) \quad (29)$$

or

$$\bar{X}_N \approx N\left(\mu, \frac{\sigma^2}{N}\right). \quad (30)$$