CSE 6740 Lecture 15
What Error Guarantees Can We Make?
(Learning Theory and Generalization)

Alexander Gray
agray@cc.gatech.edu

Georgia Institute of Technology
Today

1. Statistical inequalities (*How can we bound values that can appear in the future?*)
2. Confidence intervals (*How good is the estimation/learning?*)
Statistical inequalities

How can we bound values that can appear in the future?
Markov’s Inequality

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

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}.$$  \hfill (1)
Markov’s Inequality: Proof

Since $X > 0$,

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

(2)

$$
= \int_0^t x f(x) \, dx + \int_t^\infty x f(x) \, dx
$$

(3)

$$
\geq \int_t^\infty x f(x) \, dx
$$

(4)

$$
\geq \int_t^\infty f(x) \, dx
$$

(5)

$$
\geq t \int_t^\infty f(x) \, dx
$$

(5)

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

(6)
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,

\[ P(|X - \mu| \geq t) = P(|X - \mu|^2 \geq t^2) \leq \frac{\mathbb{E}(X - \mu)^2}{t^2} \]

(9)

\[ \leq \frac{\sigma^2}{t^2}. \]

(10)

(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 $\epsilon$ of $p$?
Chebyshev’s Inequality: Example

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

$$\mathbb{P}(|\overline{X}_N - p| > \epsilon) \leq \frac{\mathbb{V}(\overline{X}_N)}{\epsilon^2}$$  \hspace{1cm} (12)

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

$$\leq \frac{1}{4N\epsilon^2}$$ \hspace{1cm} (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 $X_1, \ldots, X_N$ 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 $X_1, \ldots, X_N \sim \text{Bernoulli}(p)$. Then for any $\epsilon > 0$,

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

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

Hoeffding’s Inequality: Example

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

$$\mathbb{P}( |\overline{X}_{100} - p| > .2 ) \leq .0625.$$  \hspace{1cm} (17)

Hoeffding’s inequality yields the tighter bound

$$\mathbb{P}( |\overline{X}_{100} - p| > .2 ) \leq 2e^{-2 \cdot 100(.2)^2} = .00067.$$  \hspace{1cm} (18)
(Weak) Law of Large Numbers

Theorem (WLLN): If \( X_1, \ldots, X_N \) 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 \( \mu \) 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,

$$\mathbb{P}(|\bar{X}_N - \mu| > \epsilon) \leq \frac{\text{V}(\bar{X}_N)}{\epsilon^2}$$  \hspace{1cm} (19)

$$= \frac{\sigma^2}{N \epsilon^2}$$  \hspace{1cm} (20)

which approaches 0 as $N \rightarrow \infty$. 
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 $\theta$ is an interval $C_N = (a, b)$ where $a = a(X_1, \ldots, X_N)$ and $b = b(X_1, \ldots, X_N)$ are functions of the data such that

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

(21)

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

If $\theta$ is a vector we have a confidence set, which is a ball or ellipse.
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

\[
\epsilon_N = \sqrt{\frac{1}{2N} \log \left( \frac{2}{\alpha} \right)}.
\]  

(22)

By Hoeffding’s inequality,

\[
P(|\bar{X}_N - p| > \epsilon_N) \leq 2e^{-2N\epsilon_N^2} = \alpha.
\]  

(23)

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

\[
\frac{\hat{\theta}_N - \theta}{\text{se}} \xrightarrow{\text{d}} \mathcal{N}(0, 1).
\]  

(24)

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

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

Let \( z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2)) \), that is, \( \mathbb{P}(Z > z_{\alpha/2} = \alpha/2 \) and \( \mathbb{P}(-z_{\alpha/2} < Z < a_{\alpha/2}) = 1 - \alpha \) 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 \( \hat{\theta}_N \pm 2\hat{se} \).

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

Theorem (CLT): If $X_1, \ldots, X_N$ are IID (with any distribution), with mean $\mu$ and variance $\sigma^2$, then

$$Z_N = \frac{\bar{X}_N - \mu}{\sqrt{\text{Var}(\bar{X}_N)}} = \frac{\sqrt{N}(\bar{X}_N - \mu)}{\sigma} \rightsquigarrow Z$$

(25)

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

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

(26)
Central Limit Theorem

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

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

(27)

or

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

(28)
VC Dimension

The previous estimates of optimism are mostly applicable to simple models and are likelihood-based. A more general measure of model complexity is given by VC theory.

The Vapnik-Chervonenkis (VC) Dimension of the function class $\mathcal{F}_\theta : \mathbb{R}^D \rightarrow \{0, 1\}$ indexed by parameter(s) $\theta$ is defined to be the largest number of points (in some configuration and under some labeling) that can be shattered by members of the class.

A set of points is said to be shattered by a class of functions if, no matter how we assign the location and label for each point, a member of the class can perfectly separate them.
VC Dimension

For example, the linear decision function can shatter any three points in the plane, but not any four. Hence the VC dimension of this function class is 3. In general the linear decision function in $D$ dimensions has VC dimension $D + 1$, which is also the number of free parameters.
VC Dimension

A nonlinear family may have infinite VC dimension, because by appropriate choice of its parameters $\theta$, any set of points can be shattered by this class.

The VC dimension can be defined for a real-valued function class $\mathcal{F}_\theta : \mathbb{R}^D \to \mathbb{R}$ as the VC dimension of the indicator class $\mathcal{F}_\theta = \{I(f_\theta(x) - \beta > 0)\}$, where $\beta$ takes values over the range of $f$. 
VC Bounds

For two-class classification, for every model in a function class with VC dimension $h$, with probability $1 - \eta$

$$R(M) \leq \hat{R}_{tr}(M) + \frac{\epsilon}{2} \left(1 + \sqrt{1 + \frac{4\hat{R}_{tr}(M)}{\epsilon}}\right)$$

where

$$\epsilon = c_1 \frac{h[\log(c_2 N/h) + 1] - \log(\eta/4)}{N}.$$  \hspace{1cm} (30)

For regression,

$$R(M) \leq \frac{\hat{R}_{tr}(M)}{(1 - c_3 \sqrt{\epsilon})^+}.$$  \hspace{1cm} (31)

‘Recommended’ values are $c_1 = c_2 = c_3 = 1$ for regression.
The bounds suggest that the optimism increases with $h$ and decreases with $N$ in qualitative agreement with the AIC term $|M|/N$.

Note that these bounds hold for every member (parameter setting) of the function class; AIC described the behavior of a specific member of the class (the MLE).

*Structural risk minimization* chooses the function class with the smallest value of the upper bound.

Note that VC theory does not give an actual estimate of the test error, only an upper bound on it. It is also not always easy to derive the VC dimension of a model class.
Main Things You Should Know

- Basic inequalities
- What the VC dimension is
Sample Final Questions

1. (T/F) The VC dimension is a property of a model with specific parameters.

2. (T/F) Markov’s inequality holds for any random variable.