CPSC 550: Machine Learning II	**2008/9 Term 2**

Lecture 3 — Jan 19, 2009

Lecturer: Nando de Freitas *Scribe: Bo Chen*

This lecture introduces the confidence interval and three types of convergence, and it proves the Weak Law of Large Numbers.

3.1 Review

- **Estimator**:
$$\hat{\theta}_n = g(X_{1:n})$$
- **Consistency**:
$$\hat{\theta}_n \xrightarrow{p} \theta$$
- **Markov's Inequality**: $\forall t > 0$
$$P(f(X) > t) \leq \frac{\mathbb{E}(f(X))}{t}$$
- **Chebyshev's Inequality**: $\forall t > 0, \mu = \mathbb{E}(X)$,
$$P((X-\mu)^2 > t^2) = P(|X-\mu| > t) \leq \frac{\mathbb{V}(X)}{t^2}$$

Note that there exist other, tighter bounds than these. For example, one can derive the **Chernoff bound** by exploring the following analysis of Markov's inequality (Ch. 4, Wasserman):

$$P(X > t) = p(e^X > e^t) \leq \frac{\mathbb{E}(e^X)}{e^t}$$

Question: Is Maximum Likelihood the correct thing to do? What are the valid ways of learning?

An example using Chebyshev's Inequality

(Chapter 4.3, p.64 of Wasserman)

Assume we have a predictor (neural network, SVM, logitBoost, ...). Let $\{X_i\}_{i=1}^n$ be a set of *i.i.d.* Bernoulli [1] random variables:

$$X_i = \begin{cases} 1 & \text{if prediction is correct} \\ 0 & \text{if prediction is incorrect} \end{cases}$$

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the observed error rate.

- What is the true error rate θ? i.e., what is the parameter of the Bernoulli distribution $\theta = P(X_i = 1)$?
- How likely is $\bar{X}_n$ to be within ϵ of true paramater θ?

Using Chebyshev's inequality:

$$P(|\bar{X}_n - \theta| > \epsilon) \leq \frac{\mathbb{V}(\bar{X}_n)}{\epsilon^2} = \frac{\mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i)}{\epsilon^2} \tag{3.1}$$

$$= \frac{1}{n^2\epsilon^2}\sum_{i=1}^{n} \mathbb{V}(X_i) \tag{3.2}$$

$$= \frac{\theta(1-\theta)}{n\epsilon^2} \tag{3.3}$$

where step 3.2 exploits the *i.i.d.* assumption.

Since $\theta(1-\theta) \leq 1/4$ always holds:

$$P(|\bar{X}_n - \theta| > \epsilon) \leq \frac{1}{4n\epsilon^2} \equiv \alpha$$

As a quantitative example, if $\epsilon = 0.2$, $n = 100$, then $\alpha = 0.0625$.

[1] Recall that Bernoulli$(x;\theta) \sim \theta^x(1-\theta)^{1-x}$, with mean θ and standard deviation $\theta(1-\theta)$.

3.2 Confidence Intervals

From the example,

$$\begin{aligned}\alpha &= \frac{1}{4n\epsilon^2}\\ \epsilon_n &= \frac{1}{2}\sqrt{\frac{1}{\alpha n}}\end{aligned}$$

Note that ϵ_n is a function of n.

Let $\mathcal{C} = (\bar{X}_n - \epsilon_n, \bar{X}_n + \epsilon_n)$; then we have:

$$\begin{aligned}P(\theta \notin \mathcal{C}) &= P(|\bar{X}_n - \theta| > \epsilon_n) \leq \alpha\\ P(\theta \in \mathcal{C}) &= 1 - P(|\bar{X}_n - \theta| > \epsilon_n) \geq 1 - \alpha\end{aligned}$$

In other words, the confidence interval $\mathcal{C}$ "traps" the true parameter θ with probability $1 - \alpha$.

3.3 Convergence

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables where $X_n \sim F_n$, and let $X \sim F$ be another random variable. We discuss senses in which X_n can be said to converge to X.

3.3.1 Convergence in Probability (Weak Convergence)

X_n converges to X **in probability**, or $X_n \xrightarrow{p} X$, if $\forall \epsilon > 0$:

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

Application: Weak Law of Large Numbers

Theorem 3.1. *Let $X_{1:n}$ be i.i.d. random variables.*

- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $\mu = \mathbb{E}(X_1) = \mathbb{E}(X_2) = \dots$

- $\sigma^2 = \mathbb{V}(X_1) = \mathbb{V}(X_2) = \dots$

Then $P(|\bar{X}_n - \mu| > \epsilon) \to 0$, *as* $n \to \infty$; *i.e.,* $\bar{X}_n \xrightarrow{p} \mu$.

Proof:

$$\begin{aligned}
P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\mathbb{V}(\bar{X}_n)}{\epsilon^2} &= \frac{\mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i)}{\epsilon^2} \\
&= \frac{1}{n\epsilon^2}\mathbb{V}(X_1) \\
&= \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \text{as } n \to \infty
\end{aligned}$$

□

3.3.2 Convergence in Distribution (Law)

We say X_n converges to X **in distribution**, or $X_n \rightsquigarrow X$, or $X_n \xrightarrow{d} X$, if $\forall a$ at which F is continuous:

$$\lim_{n\to\infty} F_n(a) = F(a)$$

3.3.3 Convergence in Quadratic Mean (L2 Norm)

X_n converges to X **in quadratic mean**, or $X_n \xrightarrow{q.m.} X$, if:

$$\lim_{n\to\infty} \mathbb{E}(X_n - X)^2 = 0$$