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

Lecture 4 — Jan 21, 2009

Lecturer: Nando de Freitas *Scribe: Sancho McCann*

Notation

- $a_n \to a$ means a_n converges to a, as $n \to \infty$.
- $X_n \xrightarrow{qm} X$ means X_n converges in quadratic mean to X.
- $X_n \xrightarrow{p} X$ means X_n converges in probability to X.
- $X_n \rightsquigarrow X$ means X_n converges in distribution to X.

Theorem (Wasserman 5.4)

a) $X_n \xrightarrow{qm} X \Rightarrow X_n \xrightarrow{p} X$

b) $X_n \xrightarrow{p} X \Rightarrow X_n \rightsquigarrow X$

"Convergence in distribution is the weakest"

Proof: a)

$$
\begin{aligned}
P(|X_n - X| > \epsilon) &\leq \frac{\mathbb{E}[(X_n - X)^2]}{\epsilon^2} \text{ (by Markov's inequality)} \\
\frac{\mathbb{E}[(X_n - X)^2]}{\epsilon^2} &\xrightarrow{qm} 0 \\
P(|X_n - X| > \epsilon) &\to 0
\end{aligned}
$$

□

Notation

estimator: $\hat{\theta}_n = \mathrm{g}(X_{1:n})$, a function of the data

sampling distribution: distribution of random variable $\hat{\theta}_n$

standard error (se): $\sqrt{\mathrm{var}(\hat{\theta}_n)}$

Mean squared error (MSE)

$$\begin{aligned} \text{MSE} &= \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 \\ &= \int [\text{g}(X_{1:n}) - \theta)^2] P(X_{1:n}, \theta) dX_{1:n} \end{aligned}$$

where $P(X_{1:n}, \theta) = \prod_{i=1}^{n} p(X_i, \theta)$ when data X_i are i.i.d.

Theorem 6.9

$\text{MSE} = \text{bias}(\hat{\theta})^2 + \text{var}_\theta(\hat{\theta})$

Proof

Let:

$$\begin{aligned} \bar{\theta}_n &= \mathbb{E}_\theta(\hat{\theta}_n) \\ &= \int \text{g}(X_{1:n}) P(dX_{1:n}; \theta) \\ &= \int \text{g}(X_{1:n}) dP(X_{1:n}; \theta) \\ &= \int \text{g}(X_{1:n}) P(X_{1:n}; \theta) dX_{1:n} \end{aligned}$$

Then:

$$\begin{aligned} \text{MSE} = \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2] &= \mathbb{E}_\theta[(\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta)^2] \\ &= \mathbb{E}_\theta[(\hat{\theta}_n - \bar{\theta}_n)^2 + 2(\hat{\theta}_n - \bar{\theta}_n)(\bar{\theta}_n - \theta) + (\bar{\theta}_n - \theta)^2] \\ &= \mathbb{E}_\theta[(\hat{\theta}_n - \bar{\theta}_n)^2] + 2\mathbb{E}_\theta[(\hat{\theta}_n - \bar{\theta}_n)(\bar{\theta}_n - \theta)] + \mathbb{E}_\theta[(\bar{\theta}_n - \theta)^2] \\ &= \mathbb{E}_\theta[(\hat{\theta}_n - \bar{\theta}_n)^2] + 2(\bar{\theta}_n - \theta)\mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}_n) + (\hat{\theta}_n - \theta)^2 \\ &= \mathbb{E}_\theta[(\hat{\theta}_n - \bar{\theta}_n)^2] + 2(\bar{\theta}_n - \theta)(0) + (\hat{\theta}_n - \theta)^2 \\ &= \mathbb{E}_\theta[(\hat{\theta}_n - \bar{\theta}_n)^2] + (\hat{\theta}_n - \theta)^2 \\ &= \text{var}_\theta(\hat{\theta}_n) + \text{bias}(\hat{\theta})^2 \end{aligned}$$

Consistency

An estimator $\hat{\theta}_n$ of parameter θ is said to be *consistent* iff $\hat{\theta}_n \xrightarrow{p} \theta$.

Theorem (Wasserman 6.10)

If bias and variance go to zero as $n \to \infty$, then $\hat{\theta}_n$ is consistent.

Proof: If the bias and variance both converge to zero, then by Theorem 6.9, it follows that the MSE converges to 0, i.e. $\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 \to 0$. Thus $\hat{\theta}_n \xrightarrow{qm} \theta$. By Theorem 5.4(a), it follows that $\hat{\theta}_n \xrightarrow{p} \theta$. QED. $\square$