The Law of Large Numbers and the Central Limit Theorem

Guy Lebanon

September 3, 2010

Theorem 1 (Markov Inequality)**.** *For a non-negative scalar RV X (with finite expectation and variance)*

$$P(X \geq a) \leq \frac{\mathsf{E}(X)}{a}, \qquad \forall a > 0$$

Proof. $\int_{-\infty}^{\infty} x f_X(x)dx = \int_0^{\infty} x f_X(x)dx \geq \int_a^{\infty} x f_X(x)dx \geq \int_a^{\infty} a f_X(x)dx = aP(X \geq a)$. (replace integrals with sums for discrete RVs) □

Theorem 2 (Chebyshev Inequality)**.** *For a scalar RV X (with finite expectation and variance)*

$$P(|X - E(X)| \geq a) \leq \frac{\mathsf{Var}(X)}{a^2}, \qquad \forall a > 0$$

Proof. Apply Markov inequality $P(Z^2 \geq a^2) \leq \frac{\mathsf{E}(Z^2)}{a^2}$ for the RV Z^2 where $Z = |X - E(X)|$ i.e., $P(|X - E(X)| \geq a) = P(Z \geq a) = P(Z^2 \geq a^2) \leq \frac{\mathsf{Var}(X)}{a^2}$. □

Definition 1. Let $Y^{(1)}, Y^{(2)}, \ldots$ be a sequence of random vectors and Z be a random vector. We say that $Y^{(n)}$ converges in probability to Z, and denote $Y^{(n)} \xrightarrow{\mathrm{P}} Z$, if

$$\lim_{n\to\infty} P(\|Y^{(n)} - Z\| \geq \epsilon) = 0, \qquad \forall \epsilon > 0.$$

Theorem 3 (The Weak Law of Large Numbers)**.** *Let $X^{(1)}, X^{(2)}, \ldots$ be a sequence of d-dimensional iid*[1] *random vectors with finite expectation vector μ and covariance matrix Σ. Then the sequence $Y^{(n)} = \bar{X} = \frac{1}{n}\sum_{i=1}^n X^{(i)}$ converges in probability to μ i.e.,*

$$\lim_{n\to\infty} P\left(\left\|\frac{1}{n}\sum_{i=1}^n X^{(i)} - \mu\right\| \geq \epsilon\right) = 0, \quad \forall \epsilon > 0.$$

Proof. For scalars RVs $X^{(i)}$ ($d = 1$) with expectation μ and variance σ^2 the proof follows from noting that $\mathsf{E}\,\bar{X} = n^{-1}\sum_{i=1}^n EX^{(i)} = (n/n)\mu = \mu$, $\mathsf{Var}\,\bar{X} = \frac{1}{n^2}\mathsf{Var}\sum_{i=1}^n X^{(i)} = \frac{1}{n^2}\sum_{i=1}^n \mathsf{Var}\,X^{(i)} = n^{-1}\sigma^2$ (since $X^{(i)}$ are iid) and applying Chebyshev inequality to $\bar{X} = \frac{1}{n}\sum_{i=1}^n X^{(i)}$

$$P\left(\left|\bar{X} - \mathsf{E}\,\bar{X}\right| \geq \epsilon\right) = P\left(|\bar{X} - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} \longrightarrow_{n\to\infty} 0.$$

For $d > 1$ we proceed by applying Boole's inequality and the one dimensional result above

$$\begin{aligned} P\left(\|\bar{X} - \mathsf{E}\,\bar{X}\| \geq \epsilon\right) &= P\left(\sum_{j=1}^d |\bar{X}_j - \mu_j|^2 \geq \epsilon^2\right) \leq \sum_{j=1}^d P\left(|\bar{X}_j - \mu_j|^2 \geq \epsilon^2/d\right) \\ &= \sum_{j=1}^d P\left(|\bar{X}_j - \mu_j| \geq \epsilon/\sqrt{d}\right) \leq \frac{d}{n\epsilon^2}\sum_{j=1}^d \mathsf{Var}\left(X_j\right) = \frac{d}{n\epsilon^2}\mathsf{trace}(\mathsf{Var}\left(X\right)) \to_{n\to\infty} 0. \end{aligned}$$

□

[1] independent and identically distributed i.e. all $X^{(i)}$ have the same distribution, and in particular same mean and variance.

The WLLN shows that the sequence of random vectors $Y^{(n)} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$, $n = 1, 2, \ldots$ are increasingly centered around the expectation vector μ and thus for large n, they become a good approximation to μ. The central limit theorem below is characterizing the distribution with which this convergence happens. Interestingly, it shows that the distribution of $\bar{X}$ for large n is Gaussian centered around μ with variance Σ/n. Thus, not only can we say that $\bar{X}$ is close to μ, we can say that its distribution is a bell shaped curve centered at μ whose width (variance) decays linearly with n.

Definition 2. We say that a sequence of random vectors $Y^{(1)}, Y^{(2)}, \ldots$ converges in distribution to a random vector Z, denoted $Y^{(n)} \rightsquigarrow Z$ if[2]

$$\lim_{n\to\infty} P(Y^{(n)} \in A) = P(Z \in A)$$

Theorem 4 (Central Limit Theorem)**.** *Let $X^{(1)}, X^{(2)}, \ldots$ be a sequence of iid d-dimensional random vectors RVs with finite expectation vector μ and covariance matrix Σ. Then $Y^{(n)} = \sqrt{n}(\bar{X} - \mu) \rightsquigarrow N(0, \Sigma)$ i.e.,*

$$\lim_{n\to\infty} P(\sqrt{n}(\bar{X} - \mu) \in A) = P(Z \in A), \qquad Z \sim N(0, \Sigma).$$

The 1-dimensional analog ($d = 1$ above) is

$$\lim_{n\to\infty} P(\sqrt{n}(\bar{X} - \mu) \in A) = P(Z \in A) = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} e^{-z^2/(2\sigma^2)} dz.$$

For large[3] n we can say that the approximation in the CLT above holds with high precision. Since *if* $\sqrt{n}(\bar{X} - \mu) \sim N(0, \Sigma)$ we have $\bar{X} - \mu \sim N(0, \Sigma/n)$ and $\bar{X} \sim N(\mu, \Sigma/n)$ (why?) we can say that *intuitively* $\bar{X}$ has a distribution that is approximately $N(\mu, \Sigma/n)$. Similarly, *intuitively*, the distribution of $\sum_{i=1}^{n} X^{(i)}$ is approximately $N(n\mu, n^2\Sigma/n) = N(n\mu, n\Sigma)$.

The CLT is very useful and surprising. It states that if we are looking for a distribution of a RV Y that is a sum of many other iid RVs $Y^{(n)} = X^{(1)} + \cdots + X^{(n)}$ we are guaranteed that its distribution will be close to normal (for large n) regardless of the distribution of the original $X^{(i)}$.

Example: In a restaurant orders $X^{(1)}, X^{(2)}, \ldots$ are received in an iid fashion with expectation \$8 and variance \$4 (note that the distribution of the orders is unknown). The owner is interested in computing the probability that sum of the first 100 orders are greater or equal to 840 $P(\sum_{i=1}^{100} X^{(i)} \geq 840)$. To approximate that quantity using the CLT we need to transform it to a form that enable us to invoke the CLT:

$$P\left(\sum_{i=1}^{100} X^{(i)} \geq 840\right) = P\left(\frac{\sum_{i=1}^{100} X^{(i)} - 8 \cdot 100}{2 \cdot 10} \geq \frac{840 - 8 \cdot 100}{2 \cdot 10}\right) \approx P(Z \geq 2) = \int_2^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

where the last expression is commonly found in tables or by numerical integration.

[2]This definition is slightly simplified in order to avoid measure theory. It should be sufficient for almost all practical cases.

[3]Statisticians generally agree that the approximation in the CLT is accurate when $n > 30$ in most cases. However, the CLT approximation can also be used effectively in many cases where $n < 30$.