Probability Theory

Manik Bhandari
Department of Computer Science
Indian Institute of Science
Bangalore, India
mbbhandarimanik2@gmail.com

Abstract

Notes of Probability Theory mainly from the book *Introduction to Probability Theory by Hoel, Port, Stone.*

1 Probability Spaces

2 Discrete Random Variables

3 Expectation of Discrete Random Variables

- *frequentist approach*. Let n be the number of trials and $N_n(x_i)$ be the number of times you observed x_i, then for large n
$$f(x_i) = \frac{N_n(x_i)}{n}.$$
- A random variable X has *finite* expectation if and only if
$$\sum_{i=0}^{i=\infty} |x_i| f(x_i) < \infty,$$
then the expectation is
$$EX = \sum_{i=0}^{i=\infty} x_i f(x_i).$$
Otherwise EX is undefined.
- Quick Expectations to remember: Binomial - np, Poisson - λ (so is the variance), Geometric - $\frac{1-p}{p}$.
- Expectation of a function (if it is finite):
$$E\phi(X) = \sum_x \phi(x) f(x)$$
- Properties

  $E(c_1X_1 + c_2X_2 + ...) = c_1EX_1 + c_2EX_2 + ...$
  $|EX| \leq E|X|$
  $X \leq Y \implies EX \leq EY$
  If $P(|X| \leq M) = 1$ then $EX \leq M$
  if X and Y are independent, $E(XY) = (EX)(EY)$. The converse is not true.
  if X is a non negative, integer valued random variable, X has a finite expectation if and only if $\sum_{x=0}^{x=\infty} P(X \geq x)$ converges. Then $EX = \sum_{x=0}^{x=\infty} P(X \geq x)$.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

3.1 Moments

1. EX^r is defined as the r^th *moment*. $E(X-\mu)^r$ is defined as the r^th *central moment.*
2. If EX^r exists then EX^k exists for $k \leq r$.
3. Mean (μ) is the first moment.
4. If X and Y have a moment of order r then $X+Y$ also have a moment of order r. And by induction $X_1 + X + 2 + ...X_n$ also has a moment of order r. (see pg 93 for proof)

Variance. If random variable X has a finite second moment, then

$$VarX = E[(X-EX)^2] \implies VarX = EX^2 - (EX)^2.$$

Variance is a measure of the spread about the mean. Consider (extreme case) when $P(X=c)=1$. $VarX = 0$ which means that X is not spread about the mean, there is no variance, it is exactly at the mean.
Consider Mean Squared Error (MSE) of a random variable X. We wish to choose an a that minimizes $MSE = E(X-a)^2$. Using calculus this value occurs at $a = EX$ and the value of MSE is $VarX$. Another way to see this is that $E(X-a)^2 = E(X-\mu)^2 + (\mu-a)^2$ which has a minima at $\mu = a$ and the min. value is the variance.

finding variance. An easy way to find variance is to use probability generating functions.

$$\phi_X(t) = \sum_{x=0}^{x=\infty} f_X(x)t^x$$

$$\implies \phi'_X(1) = EX, \phi''_X(1) = EX(X-1) \implies VarX = \phi''_X(1) - (\phi'_X(1))^2 + \phi'_X(1)$$

Standard deviation. $\sigma = \sqrt{VarX}$

Covariance.

1. $VarX+Y = VarX + VarY + 2E[(X-EX)(Y-EY)]$
2. $Cov(X,Y) = E[(X-EX)(Y-EY)] = E(XY) - (EX)(EY)$
3. So $VarX+Y = VarX + VarY + 2Cov(X,Y)$
4. $Cov(X,Y) = 0$ for independent X, Y. (converse is not true)
5. *using induction*

$$Var(\sum_{i=0}^{i=n} X_i) = \sum_{i=0}^{i=n} VarX_i + 2\sum_{i=0}^{i=n-1}\sum_{j=1}^{j=n} Cov(X_i, X_j).$$

If independent, $Var(\sum_{i=0}^{i=n} X_i) = \sum_{i=0}^{i=n} VarX_i$. And if the variance is common, $Var(\sum_{i=0}^{i=n} X_i) = n\sigma^2$

Correlation coefficient. Measures the degree of dependence.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{(VarX)(VarY)}}.$$

If X, Y are independent, $\rho = 0$ (since $Cov(X,Y) = 0$). ρ is always between -1 and 1 and $|\rho| = 1$ only when $P(X = aY) = 1$ (highly dependent).

Theorem 1. *The schwarz inequality. Let X and Y have finite second moment. Then*

$$E(XY)^2 \leq (EX)^2(EY)^2.$$

The equality holds if and only if $P(Y=0) = 1$ *or* $P(X = aY) = 1$ *for some constant a.*

Proof. If $P(Y = 0) = 1$, equality holds and $P(X = aY) = 1$, equality holds. Let $P(Y = 0) < 1$. $\implies EY^2 > 0$.
Consider $E(X - \lambda Y)^2$. Clearly, $0 \leq E(X - \lambda Y)^2 = \lambda^2 EY^2 - 2\lambda EXY + EX^2$ which is a quadratic in λ. It has a minima at $\lambda = \frac{E(XY)}{EY^2}$. So the minimum value is positive and is given by

$$E(X - \lambda Y)^2 = EX^2 - \frac{[E(XY)]^2}{EY^2} \geq 0.$$

If equality holds then $E[x - \lambda Y] = 0 \implies P(X - \lambda Y = 0) = 1 \implies P(X = \lambda Y) = 1.$ $\square$

Chebyshev's Inequality. Let X be a non-negative random variable having finite expectation. Let t be a real number, define Y such that $Y = 0$ if $X < t$ and $Y = 1$ if $X >= t$. $EY = 0P(Y = 0) + tP(Y = t) = tP(X \geq t)$. Since $X \geq Y, EX \geq EY = tP(X \geq t)$

$$\implies P(X \geq t) \leq \frac{EX}{t}.$$

Thus

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

This assumes nothing about the distribution of X. Usually you get better bounds if you know something about the distribution of X (apart from it being nonnegative and having finite variance).

Application. Weak Law of Large numbers Let S_n be sum of n random variables having common variance σ^2. Then $Var(S_n/n) = n\sigma^2/n^2 = \sigma^2/n$. So

$$P(|\frac{S_n}{n} - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2} \implies \lim_{n\to\infty} P(|\frac{S_n}{n} - \mu| \geq \delta) = 0$$

for any $\delta > 0$. Which means that you can approximate μ by S_n/n for large n.