12/8/25

Steven Karp - Substitute Lecturer

Law of large numbers (Theorem 6.1)

(informally:) If you flip a fair coin many many times, then the proportion of heads will be very very close to $\frac{1}{2}$

Let $X_1, \dots, X_n$ be i.i.d. (independent and indentically distributed) random variables with mean μ and variance σ^2. Consider the sum:

$$S_n = X_1 + \dots + X_n \qquad (E[S_n] = \mu n)$$

(e.g. $X_i \sim Ber(\frac{1}{2})$, S_n = # heads after n flips)

For $\varepsilon > 0$: $$P(\underbrace{\left|\frac{S_n}{n} - \mu\right| > \varepsilon}_{\text{error} > \varepsilon}) \leq \frac{\sigma^2}{n\varepsilon^2}$$

e.g. Suppose we flip a coin 100 times and get 20 heads.

$$\frac{S_{100}}{100} = \frac{20}{100} = \frac{1}{5}, \quad \text{so} \quad \left|\frac{S_n}{n} - \mu\right| = \left|\frac{1}{5} - \frac{1}{2}\right| = \frac{3}{10}$$

$$P(\text{error} > \tfrac{3}{10}) \leq \frac{\sigma^2}{n\varepsilon^2} = \frac{\frac{1}{2}(1-\frac{1}{2})}{100(\frac{3}{10})^2} = \frac{1}{36}$$

The actual probability is much lower than $\frac{1}{36}$ (just an upper bound)

Central Limit Theorem

Same setup. We want to know how S_n behaves.

$$E[S_n] = n\mu$$

$$\operatorname{var}(S_n) = \operatorname{var}(X_1 + \dots + X_n)$$

since $X_1, \dots, X_n$ are independent. (hence pairwise uncorrelated)

$$= \operatorname{var}(X_1) + \dots + \operatorname{var}(X_n) = n\sigma^2$$

The Central Limit Theorem says that for very large n,

$$S_n \overset{\text{(approx.)}}{\approx} N(n\mu, n\sigma^2) \leftarrow \text{Gaussian / Normal}$$

There are various precise versions of the CLT.

The classical CLT (Theorem 6.5) says that the sequence of random variables $\left(\frac{S_n - n\mu}{\sqrt{n}\,\sigma}\right)_{n=1}^{\infty}$ converges <u>in-distribution</u> to $N(0,1)$, i.e. the cdfs converge pointwise to the cdf of $N(0,1)$.

Ex: Suppose customers of a restaurant spend on average \$14 per visit with a standard deviation of \$9.

suppose there are 100 customers today. What is the probability that the total revenue today exceeds \$1300

By the CLT we can approximate total revenue by

$$N(n\mu, n\sigma^2) = N(100 \cdot 14,\ 100 \cdot 9^2)$$

$$= N(1400,\ 90^2)$$

So we can approximate the probability exceeds \$1300

by $P(N(1400, 90^2) > 1300)$

$$= 1 - P(N(1400, 90^2) \leq 1300)$$

$\approx 86.7\%$ (calculated using R program)

Mean-squared error (MSE)

Suppose we have a model which uses X to predict Y.

(imagine we have two stocks), and our model is that

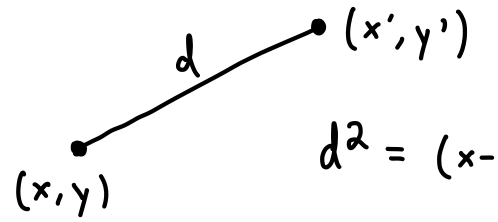
$Y \approx f(X)$. The MSE of the approximation is

$$E\left[(Y - f(X))^2\right]$$

(this is also squared error in the sense of least squares)

Why use squared error? Two main reasons:

(i) It is nice linear-algebraically (it comes from the squared euclidian distance)



$$d^2 = (x - x')^2 + (y - y')^2$$

(ii) CLT.

↳ Suppose Y is exactly a linear function of X

plus some noise: $Y = \underbrace{\alpha + \beta X}_{\text{linear}} + \underbrace{Z}_{\text{noise}}$

($Z \sim \underbrace{N(0, \sigma^2)}_{\text{why? CLT. (otherwise mean would just contribute to } \alpha)}$

Now suppose we have n data points

$(x_1, y_1), \ldots, (x_n, y_n)$ sampled from (X, Y)

and we want to find the "best" model (best α, β)

→ one notion of best is the maximum likelihood estimator (MLE), which maximizes the probability (density) of observing the given data from our model. Let $p(x)$ be the pdf of $N(0, \sigma^2)$. The probability density of observing the given data is:

$$\prod_{i=1}^{n} p(y_i - (\alpha + \beta x_i))$$

!!!

$$= \prod_{i=1}^{n} c_1 \exp(-c_2 (y_i - (\alpha + \beta x_i))^2)$$

$$= c_1^n \exp\Big(-c_2 \underbrace{\Big[\sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2\Big]}_{\text{want to minimize this (just the sample MSE!)}}\Big)$$