CS70: Jean Walrand: Lecture 31.

Nonlinear Regression

1. Review: joint distribution, LLSE
2. Quadratic Regression
3. Definition of Conditional expectation
4. Properties of CE
5. Applications: Diluting, Mixing, Rumors
6. CE = MMSE

Review

Definitions Let X and Y be RVs on Ω.

- Joint Distribution: $Pr[X=x, Y=y]$
- Marginal Distribution: $Pr[X=x] = \sum_y Pr[X=x, Y=y]$
- Conditional Distribution: $Pr[Y=y|X=x] = \frac{Pr[X=x,Y=y]}{Pr[X=x]}$
- LLSE: $L[Y|X] = a+bX$ where a,b minimize $E[(Y-a-bX)^2]$.

We saw that

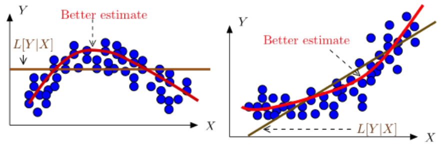
$$L[Y|X] = E[Y] + \frac{cov(X,Y)}{var[X]}(X - E[X]).$$

Recall the non-Bayesian and Bayesian viewpoints.

Nonlinear Regression: Motivation

There are many situations where a good guess about Y given X is not linear.

E.g., (diameter of object, weight), (school years, income), (PSA level, cancer risk).



Our goal: explore estimates $\hat{Y} = g(X)$ for nonlinear functions $g(\cdot)$.

Quadratic Regression

Let X, Y be two random variables defined on the same probability space.

Definition: The quadratic regression of Y over X is the random variable

$$Q[Y|X] = a + bX + cX^2$$

where a,b,c are chosen to minimize $E[(Y-a-bX-cX^2)^2]$.

Derivation: We set to zero the derivatives w.r.t. a,b,c. We get

$$\begin{aligned} 0 &= E[Y-a-bX-cX^2] \\ 0 &= E[(Y-a-bX-cX^2)X] \\ 0 &= E[(Y-a-bX-cX^2)X^2] \end{aligned}$$

We solve these three equations in the three unknowns (a,b,c).

Note: These equations imply that $E[(Y-Q[Y|X])h(X)] = 0$ for any $h(X) = d + eX + fX^2$. That is, the estimation error is orthogonal to all the quadratic functions of X. Hence, $Q[Y|X]$ is the projection of Y onto the space of quadratic functions of X.

Conditional Expectation

Definition Let X and Y be RVs on Ω. The conditional expectation of Y given X is defined as

$$E[Y|X] = g(X)$$

where

$$g(x) := E[Y|X=x] := \sum_y yPr[Y=y|X=x].$$

Fact

$$E[Y|X=x] = \sum_\omega Y(\omega)Pr[\omega|X=x].$$

Proof: $E[Y|X=x] = E[Y|A]$ with $A = \{\omega : X(\omega) = x\}$. $\square$

Deja vu, all over again?

Have we seen this before? Yes.

Is anything new? Yes.

The idea of defining $g(x) = E[Y|X=x]$ and then $E[Y|X] = g(X)$.

Big deal? Quite! Simple but most convenient.

Recall that $L[Y|X] = a + bX$ is a function of X.

This is similar: $E[Y|X] = g(X)$ for some function $g(\cdot)$.

In general, $g(X)$ is not linear, i.e., not $a+bX$. It could be that $g(X) = a+bX+cX^2$. Or that $g(X) = 2\sin(4X) + \exp\{-3X\}$. Or something else.

Properties of CE

$$E[Y|X=x] = \sum_y yPr[Y=y|X=x]$$

Theorem
(a) X, Y independent $\Rightarrow E[Y|X] = E[Y]$;
(b) $E[aY+bZ|X] = aE[Y|X] + bE[Z|X]$;
(c) $E[Yh(X)|X] = h(X)E[Y|X], \forall h(\cdot)$;
(d) $E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot)$;
(e) $E[E[Y|X]] = E[Y]$.

Proof:

(a),(b) Obvious
(c) $E[Yh(X)|X=x] = \sum_\omega Y(\omega)h(X(\omega)Pr[\omega|X=x]$
$$= \sum_\omega Y(\omega)h(x)Pr[\omega|X=x] = h(x)E[Y|X=x]$$

Properties of CE

$$E[Y|X=x] = \sum_y yPr[Y=y|X=x]$$

Theorem
(a) X, Y independent $\Rightarrow E[Y|X] = E[Y]$;
(b) $E[aY+bZ|X] = aE[Y|X] + bE[Z|X]$;
(c) $E[Yh(X)|X] = h(X)E[Y|X], \forall h(\cdot)$;
(d) $E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot)$;
(e) $E[E[Y|X]] = E[Y]$.

Proof: (continued)

(d)
$$\begin{aligned} E[h(X)E[Y|X]] &= \sum_x h(x)E[Y|X=x]Pr[X=x] \\ &= \sum_x h(x)\sum_y yPr[Y=y|X=x]Pr[X=x] \\ &= \sum_x h(x)\sum_y yPr[X=x,y=y] \\ &= \sum_{x,y} h(x)yPr[X=x,y=y] = E[h(X)Y]. \end{aligned}$$

Properties of CE

$$E[Y|X=x] = \sum_y yPr[Y=y|X=x]$$

Theorem
(a) X, Y independent $\Rightarrow E[Y|X] = E[Y]$;
(b) $E[aY+bZ|X] = aE[Y|X] + bE[Z|X]$;
(c) $E[Yh(X)|X] = h(X)E[Y|X], \forall h(\cdot)$;
(d) $E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot)$;
(e) $E[E[Y|X]] = E[Y]$.

Proof: (continued) (e) Let $h(X) = 1$ in (d).

□

Properties of CE

Theorem
(a) X, Y independent $\Rightarrow E[Y|X] = E[Y]$;
(b) $E[aY+bZ|X] = aE[Y|X] + bE[Z|X]$;
(c) $E[Yh(X)|X] = h(X)E[Y|X], \forall h(\cdot)$;
(d) $E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot)$;
(e) $E[E[Y|X]] = E[Y]$.

Note that (d) says that

$$E[(Y - E[Y|X])h(X)] = 0.$$

We say that the estimation error $Y - E[Y|X]$ is orthogonal to every function $h(X)$ of X.

We call this the projection property. More about this later.

Application: Calculating $E[Y|X]$

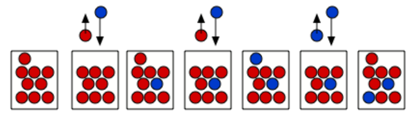
Let X, Y, Z be i.i.d. with mean 0 and variance 1. We want to calculate

$$E[2+5X+7XY+11X^2+13X^3Z^2|X].$$

We find

$$\begin{aligned} &E[2+5X+7XY+11X^2+13X^3Z^2|X] \\ &= 2+5X+7XE[Y|X]+11X^2+13X^3E[Z^2|X] \\ &= 2+5X+7XE[Y]+11X^2+13X^3E[Z^2] \\ &= 2+5X+11X^2+13X^3(var[Z]+E[Z]^2) \\ &= 2+5X+11X^2+13X^3. \end{aligned}$$

Application: Diluting



$X_1 = N$ red balls $\quad X_2 = N-1 \quad X_3 = N-2 \quad X_4 = N-2$

At each step, pick a ball from a well-mixed urn. Replace it with a blue ball. Let X_n be the number of red balls in the urn at step n. What is $E[X_n]$?

Given $X_n = m$, $X_{n+1} = m-1$ w.p. m/N (if you pick a red ball) and $X_{n+1} = m$ otherwise. Hence,

$$E[X_{n+1}|X_n = m] = m - (m/N) = m(N-1)/N = X_n\rho,$$

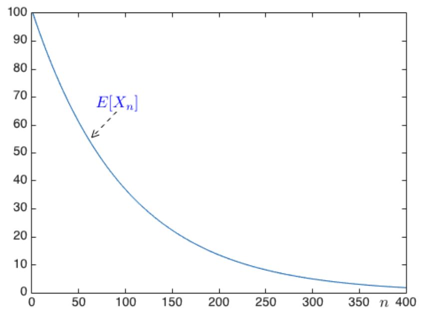
with $\rho := (N-1)/N$. Consequently,

$$E[X_{n+1}] = E[E[X_{n+1}|X_n]] = \rho E[X_n], n \geq 1.$$

$$\implies E[X_n] = \rho^{n-1}E[X_1] = N(\frac{N-1}{N})^{n-1}, n \geq 1.$$

Diluting

Here is a plot:



Diluting

By analyzing $E[X_{n+1}|X_n]$, we found that $E[X_n] = N(\frac{N-1}{N})^{n-1}, n \geq 1$.

Here is another argument for that result.

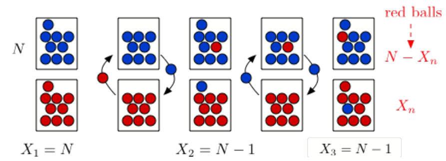
Consider one particular red ball, say ball k. At each step, it remains red w.p. $(N-1)/N$, when another ball is picked. Thus, the probability that it is still red at step n is $[(N-1)/N]^{n-1}$. Let

$$Y_n(k) = 1\{\text{ball } k \text{ is red at step } n\}.$$

Then, $X_n = Y_n(1) + \cdots + Y_n(N)$. Hence,

$$\begin{aligned} E[X_n] &= E[Y_n(1) + \cdots + Y_n(N)] = NE[Y_n(1)] \\ &= NPr[Y_n(1) = 1] = N[(N-1)/N]^{n-1}. \end{aligned}$$

Application: Mixing



At each step, pick a ball from each well-mixed urn. We transfer them to the other urn. Let X_n be the number of red balls in the bottom urn at step n. What is $E[X_n]$?

Given $X_n = m$, $X_{n+1} = m+1$ w.p. p and $X_{n+1} = m-1$ w.p. q

where $p = (1 - m/N)^2$ (B goes up, R down) and $q = (m/N)^2$ (R goes up, B down).

Thus,
$E[X_{n+1}|X_n] = X_n + p - q = X_n + 1 - 2X_n/N = 1 + \rho X_n, \ \rho := (1 - 2/N)$.

Mixing

We saw that $E[X_{n+1}|X_n] = 1 + \rho X_n, \ \rho := (1 - 2/N)$. Hence,

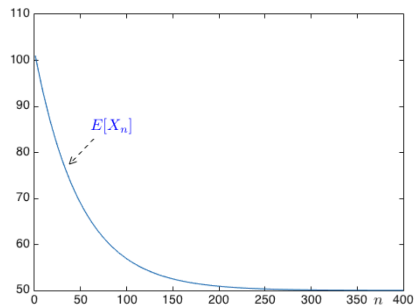
$$\begin{aligned} &E[X_{n+1}] = 1 + \rho E[X_n] \\ &E[X_2] = 1 + \rho N; E[X_3] = 1 + \rho(1 + \rho N) = 1 + \rho + \rho^2 N \\ &E[X_4] = 1 + \rho(1 + \rho + \rho^2 N) = 1 + \rho + \rho^2 + \rho^3 N \\ &E[X_n] = 1 + \rho + \cdots + \rho^{n-2} + \rho^{n-1} N. \end{aligned}$$

Hence,

$$E[X_n] = \frac{1 - \rho^{n-1}}{1 - \rho} + \rho^{n-1} N, n \geq 1.$$

Application: Mixing

Here is the plot.



Application: Going Viral

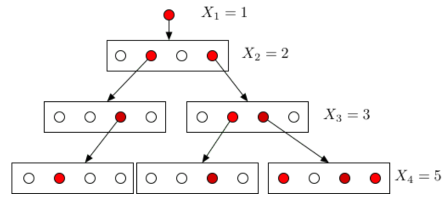
Consider a social network (e.g., Twitter).

You start a rumor (e.g., Walrand is really weird).

You have d friends. Each of your friend retweets w.p. p.

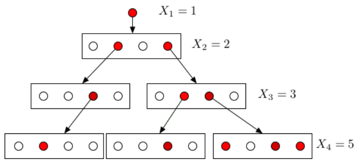
Each of your friends has d friends, etc.

Does the rumor spread? Does it die out (mercifully)?



In this example, $d = 4$.

Application: Going Viral



Fact: Let $X = \sum_{n=1}^{\infty} X_n$. Then, $E[X] < \infty$ iff $pd < 1$.

Proof:
Given $X_n = k$, $X_{n+1} = B(kd, p)$. Hence, $E[X_{n+1}|X_n = k] = kpd$.

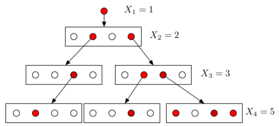
Thus, $E[X_{n+1}|X_n] = pdX_n$. Consequently, $E[X_n] = (pd)^{n-1}, n \geq 1$.

If $pd < 1$, then $E[X_1 + \cdots + X_n] \leq (1-pd)^{-1} \implies E[X] \leq (1-pd)^{-1}$.

If $pd \geq 1$, then for all C one can find n s.t.
$E[X] \geq E[X_1 + \cdots + X_n] \geq C$. $\square$

In fact, one can show that $pd \geq 1 \implies Pr[X = \infty] > 0$.

Application: Going Viral



An easy extension: Assume that everyone has an independent number D_i of friends with $E[D_i] = d$. Then, the same fact holds.

To see this, note that given $X_n = k$, and given the numbers of friends $D_1 = d_1, \ldots, D_k = d_k$ of these X_n people, one has $X_{n+1} = B(d_1 + \cdots + d_k, p)$. Hence,

$$E[X_{n+1}|X_n = k, D_1 = d_1, \ldots, D_k = d_k] = p(d_1 + \cdots + d_k).$$

Thus, $E[X_{n+1}|X_n = k, D_1, \ldots, D_k] = p(D_1 + \cdots + D_k)$.

Consequently, $E[X_{n+1}|X_n = k] = E[p(D_1 + \cdots + D_k)] = pdk$.

Finally, $E[X_{n+1}|X_n] = pdX_n$, and $E[X_{n+1}] = pdE[X_n]$.

We conclude as before.

Application: Wald's Identity

Here is an extension of an identity we used in the last slide.

Theorem Wald's Identity

Assume that $X_1, X_2, \ldots$ and Z are independent, where

Z takes values in $\{0, 1, 2, \ldots\}$

and $E[X_n] = \mu$ for all $n \geq 1$.

Then,

$$E[X_1 + \cdots + X_Z] = \mu E[Z].$$

Proof:

$E[X_1 + \cdots + X_Z | Z = k] = \mu k$.

Thus, $E[X_1 + \cdots + X_Z | Z] = \mu Z$.

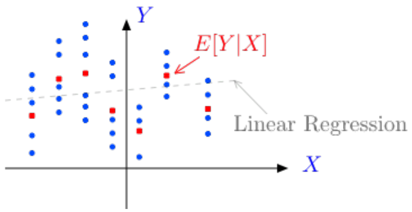
Hence, $E[X_1 + \cdots + X_Z] = E[\mu Z] = \mu E[Z]$. $\square$

CE = MMSE

Theorem
$E[Y|X]$ is the 'best' guess about Y based on X.

Specifically, it is the function $g(X)$ of X that

minimizes $E[(Y - g(X))^2]$.



CE = MMSE

Theorem CE = MMSE

$g(X) := E[Y|X]$ is the function of X that minimizes $E[(Y - g(X))^2]$.

Proof:
Let $h(X)$ be any function of X. Then

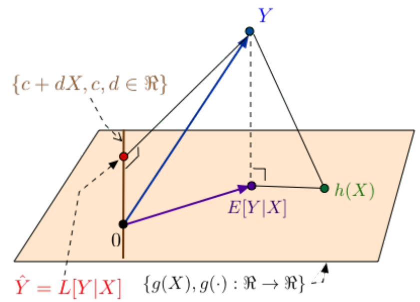
$$\begin{aligned} E[(Y - h(X))^2] &= E[(Y - g(X) + g(X) - h(X))^2] \\ &= E[(Y - g(X))^2] + E[(g(X) - h(X))^2] \\ &\quad + 2E[(Y - g(X))(g(X) - h(X))]. \end{aligned}$$

But,

$$E[(Y - g(X))(g(X) - h(X))] = 0 \text{ by the projection property.}$$

Thus, $E[(Y - h(X))^2] \geq E[(Y - g(X))^2]$. $\square$

$E[Y|X]$ and $L[Y|X]$ as projections



$L[Y|X]$ is the projection of Y on $\{a + bX, a, b \in \Re\}$: LLSE

$E[Y|X]$ is the projection of Y on $\{g(X), g(\cdot) : \Re \to \Re\}$: MMSE.

Summary

Conditional Expectation

- Definition: $E[Y|X] := \sum_y y Pr[Y = y|X = x]$
- Properties: Linearity,
  $Y - E[Y|X] \perp h(X)$; $E[E[Y|X]] = E[Y]$
- Some Applications:
  - Calculating $E[Y|X]$
  - Diluting
  - Mixing
  - Rumors
  - Wald
- MMSE: $E[Y|X]$ minimizes $E[(Y - g(X))^2]$ over all $g(\cdot)$