Econ 2120 Section 2 - Conditional Expectation

Yibo Sun and Alice Wu

13 September 2021

Outline

Review of Probability

- Probability Space
- Conditional Probability, Bayes Rule
- Expectation
- Law of Iterated Expectations

Optimal Prediction

- Regression Function (CEF) and its Properties
- $E[Y|X]$ vs. $\mathrm{E}^*[Y|X]$

Outline

Review of Probability

- Probability Space
- Conditional Probability, Bayes Rule
- Expectation
- Law of Iterated Expectations

Optimal Prediction

- Regression Function (CEF) and its Properties
- $E[Y|X]$ vs. $\mathrm{E}^*[Y|X]$

Probability Space

- The states of nature $S = \{s_1, ..., s_M\}$ contain points called elementary events.
- $A \subset S$ is an event, and let $\mathcal{B}$ denote the set of events $\sim 2^S$
- A probability measure $P : \mathcal{B} \to [0, 1]$ such that $P(s_j) \geq 0$ and $\sum_{j=1}^{M} P(s_j) = 1$, and $P(A) = \sum_{s \in A} P(s)$
- A probability space consists of the triple $(S, \mathcal{B}, P)$.

Conditional Probability

Consider an event B with $P(B) > 0$. $\forall s \in S$:

- If $s \notin B$, $P(s|B) = 0$.
- If $s \in B$, $P(s|B) = \dfrac{P(s)}{P(B)} = \dfrac{P(s)}{\sum_{s' \in B} P(s')}$

For any event A and B, the probability that both A, B occur is

$$P(A \cap B) = \sum_{s \in A \cap B} P(s)$$

the probability of A conditional on that B occurs is:

$$\begin{aligned} P(A|B) &= \sum_{s \in A} P(s|B) = \sum_{s \in A \cap B} P(s|B) \\ &= \frac{\sum_{s \in A \cap B} P(s)}{P(B)} = \frac{P(A \cap B)}{P(B)} \end{aligned}$$

Theorems

Theorem (Law of Total Probability)

If $\{B_j\}$ *forms a partition of* S ($\cup_j B_j = S,\ B_j \cap B'_j = \emptyset, \forall j \neq j'$), *then for any event* $A \subset S$, *we have:*

$$P(A) = \sum_j P(A \cap B_j) = \sum_j P(A|B_j)P(B_j)$$

Theorem (Bayes' Rule)

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_j P(A|B_j)P(B_j)}$$

where the second equality follows from the law of total probability.

Prove Bayes' Rule

Use the law of total probability to show

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Prove Bayes' Rule

Use the law of total probability to show

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Pf: Note $S = A \cup A^c$, by the law of total probability,

$$\begin{aligned} P(A \cap B) &= P(A \cap B|A) * P(A) + \underbrace{P(A \cap B|A^c) * P(A^c)}_{=0} \\ &= P(A \cap B|A) * P(A) \\ &= P(B|A)P(A) \end{aligned}$$

Likewise, can show $P(A \cap B) = P(A|B)P(B)$. Therefore, $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$.

Joint Distribution

For a pair of random variables (X,Y), we have

- A partition of state space S: all possible values of (x,y) in the Cartesian product $\mathcal{X} \times \mathcal{Y} = \{x \in \{x_1, ..., x_K\}, y \in \{y_1, ..., y_L\}\}$
- Probability measure: the joint distribution of (X,Y) given by

$$P_{XY}(x,y) = P(X=x, Y=y) = \sum_{s \in S: X(s)=x, Y(s)=y} P(s)$$

for $(x,y) \in \mathcal{X} \times \mathcal{Y}$ Single Variable

Marginal distribution for X and Y:

$$P(X) : \mathcal{X} \to [0,1], P_X(x) = \sum_{y \in \mathcal{Y}} P_{XY}(x,y)$$

$$P(Y) : \mathcal{Y} \to [0,1], P_Y(y) = \sum_{x \in \mathcal{X}} P_{XY}(x,y)$$

Conditional Distribution

Assume $P(X = x) \neq 0$, the distribution of Y conditional on X is:

$$\begin{aligned} P_{Y|X}(y|x) &= P(Y = y|X = x) \\ &= \frac{P(Y = y, X = x)}{P(X = x)} \\ &= \frac{P_{YX}(y, x)}{P_X(x)} \end{aligned}$$

which is equivalent to

$$P_{XY}(x, y) = P_{Y|X}(y|x)P_X(x)$$

Continuous variables

Let X be a continuous variable with p.d.f. f_X s.t. $\int_{-\infty}^{\infty} f_X(x) = 1$ and $Pr(X \in A) = \int_A f_X(x)dx$, for any set A.

Joint: given continuous random variables X, Y, the joint p.d.f. $f(x, y)$ satisifies

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$$

$$Pr(X \in A, Y \in B) = \int_A \int_B f(x, y)dxdy$$

Marginal:
$f_X(x) = \int_y f(x, y)dy$ and $f_Y(y) = \int_x f(x, y)dx$

Conditional:
$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$
Bayes' Rule:

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

Expectation

A scalar random variable Y is a mapping from S to a real number in $\mathbb{R}$:

$$Y : S \to \mathbb{R}$$

There are two equivalent ways to calculate the expected value of Y.

1. Sum Y's value at each state and weight by the probability of each state:

$$\mathbb{E}[Y] = \sum_{j=1}^{M} Y(s_j)P(s_j)$$

2. Sum distinct values that Y can take, weighted by the probability of each value :

$$\mathbb{E}[Y] = \sum_{l=1}^{L} y_l P(Y = y_l)$$

where $P(Y = y_l) = \sum_{j:Y(s_j)=y_l} P(s_j)$

Conditional Expectation

The expectation of Y conditional on event B happening is:

$$\mathbb{E}[Y|B] = \sum_j Y(s_j)P(s_j|B)$$

Let X be another random variable: $X : S \to \mathbb{R}$ taking values from $\{x_1, ..., x_K\}$.
Define event $B = \{s : X(s) = x_k\}$. Then

$$\begin{aligned}\mathbb{E}[Y|B] &= \sum_j Y(s_j)P(s_j|B) \\ &= \sum_j Y(s_j)P(s_j|X = x_k) \\ &= \mathbb{E}[Y|X = x_k]\end{aligned}$$

Note: $P(s_j|X = x_k) = \dfrac{P(s_j)}{P(X = x_k)}$ if $X(s_j) = x_k$. Otherwise, it's zero.

Law of Iterated Expectations (LIE)

Theorem (Law of Total / Iterated Expectations)

If Y is a random variable, for any random variable X on the same probability space,

$$E[Y] = E[E[Y|X]]$$

where the outer expectation is over X, and the inner is over $Y|X$

Law of Iterated Expectations (LIE)

Theorem (Law of Total / Iterated Expectations)

If Y is a random variable, for any random variable X on the same probability space,

$$E[Y] = E[E[Y|X]]$$

where the outer expectation is over X, and the inner is over $Y|X$

Pf 1 (Discrete):

$$\begin{aligned}
E[Y] &= \sum_y y * Pr(Y = y) \\
&= \sum_y y * (\sum_x Pr(Y = y|X = x)Pr(X = x)) \\
&= \sum_x \underbrace{(\sum_y y * Pr(Y = y|X = x))}_{E[Y|X=x]} Pr(X = x) \\
&= E_X[E[Y|X]]
\end{aligned}$$

Law of Iterated Expectations (LIE)

Pf 2 (Continuous):

$$\begin{aligned} E[Y] &= \int_y y * f(y)dy \\ &= \int_y y * (\int_x f(y|x)f(x)dx)dy \\ &= \int_x (\int_y y * f(y|x)dy)f(x)dx \\ &= E_X[E[Y|X]] \end{aligned}$$

Outline

Review of Probability

- Probability Space
- Conditional Probability, Bayes Rule
- Expectation
- Law of Iterated Expectations

Optimal Prediction

- Regression Function (CEF) and its Properties
- $E[Y|X]$ vs. $\mathrm{E}^*[Y|X]$

Regression Function (CEF)

Definition

A (mean) regression function $r : X \to \mathbb{R}$ estimates the conditional expectation of the dependent variable given the independent variable:

$$r(x) = \mathbb{E}[Y|X = x]$$

Prove:

$$\mathbb{E}[Y|X] = argmin_g \mathbb{E}[(Y - g(X))^2]$$

i.e. the CEF is the best predictor under the mean squared loss function.

Properties

Let $\epsilon = Y - \mathbb{E}[Y|X]$ Prove the following:

1. $\mathbb{E}[\epsilon] = 0$
2. $\mathbb{E}[\epsilon|X] = 0$
3. given any function h of X, $\mathbb{E}[\epsilon h(X)] = 0$

Note 3) is equivalent to say $E[Y|X]$ is the orthogonal projection of Y onto the space of *functions of* X (not just linear functions)

Properties

Pf: (1)

$$\begin{aligned} E[\epsilon] &= E[(E[Y|X] - Y)] \\ &= E_X[E[Y|X]] - E[Y] \\ &= E[Y] - E[Y] = 0 \end{aligned}$$

(2)

$$\begin{aligned} E[\epsilon|X] &= E[(E[Y|X] - Y)\,|X] \\ &= E[Y|X] - E[Y|X] = 0 \end{aligned}$$

(3)

$$\begin{aligned} E[\epsilon h(X)] &= E_X[E[\epsilon h(X)|X]] \text{ by L.I.E.} \\ &= E_X[h(X)\,\underbrace{E[\epsilon|X]}_{=0}] = 0 \end{aligned}$$

Proof for Regression Function (CEF)

Prove: $\mathbb{E}[Y|X] = argmin_g \mathbb{E}[Y - g(X)]^2$, i.e. the CEF is the best predictor under the squared loss function.
Pf:
For any function $g : X \to \mathbb{R}$,

$$
\begin{aligned}
E[(Y - g(X))^2] =& E[(Y - E[Y|X] + E[Y|X] - g(X))^2] \\
=& E[(Y - E[Y|X])^2] + E[(Y|X] - g(X))^2] \\
& + 2\underbrace{E[(Y - E[Y|X])(E[(Y|X] - g(X))]}_{E[\epsilon h(X)]=0} \\
=& E[\epsilon^2] + \underbrace{E[(Y|X] - g(X))^2]}_{\geq 0}
\end{aligned}
$$

That is, we have shown $E[(Y - E[Y|X])^2] \leq E[(Y - g(X))^2]$ for any function g.

$E[Y|X]$ vs. $\mathrm{E}^*[Y|X]$

Recall $E^*[Y|X] = X\beta$ where $\beta = argmin_b E[(Y - Xb)^2]$ is the best *linear* predictor for Y.

We have $E[Y|X] = E^*[Y|X]$ iff Y is linear in X, which would be the case if

- X is discrete: define $\delta_x = 1[X = x]$
- (Y, X) are jointly normal (uncorrelated$\sim$ independent; we will discuss this in normal linear model)

Polynomial Approximation

Theorem (Stone-Weierstrass Theorem)

If f is a continuous complex function in $[a, b]$, there exists a sequence of polynomials P_n such that

$$lim_{n\to\infty} P_n(x) = f(x)$$

uniformly on [a,b].
$(\forall \epsilon > 0 \, \exists N \;\; s.t. \;\; n > N \to \forall x \; |P_n(x) - f(x)| < \epsilon,$

This implies we can use a linear function of $\{1, X, X^2, X^3, ...\}$ to approximate the regression function

$$E[Y|X] = lim_{n\to\infty} E^*[Y|1, X, X^2, X^3, ...X^n]$$

Summary

- Regression function is the orthogonal projection onto all functions of X:

$$E[(Y - r(X))\, g(X)] = 0 \text{ for all } g$$

- Linear predictor is the orthogonal projection onto $span(X) \sim$ linear functions of X:

$$E[(Y - E^*(Y \mid 1, X))\, (\beta_0 + \beta_1 X)] = 0 \text{ for all } \beta_0, \beta_1$$

- Show $E^*(Y|1, X) = E^*(r(X)|1, X)$

Summary

Recap:

$$
\begin{aligned}
r(X) = E[Y|X] &= \arg\min_{f(X)} ||Y - f(X)||^2 \\
&= \lim_{M\to\infty} E^* \left(Y \mid 1, X, X^2, \ldots, X^M \right)
\end{aligned}
$$

$$
\begin{aligned}
E^*[Y|X] =& X\beta \\
\text{where } \beta &= \arg\min_{\beta} ||Y - \beta' X||^2 \\
&= \arg\min_{\beta} ||r(X) - \beta' X||^2
\end{aligned}
$$