7.6 조건부 기댓값과 예측문제

확률변수 Y의 기댓값을 구할때 주변 확률밀도함수 $p_Y(y)$를 사용하여 가중치를 계산하지 않고 조건부 확률밀도함수 $p_{Y|X}(y|x)$를 이용하여 가중치를 계산하면 조건부 기댓값(conditional expectation) 혹은 조건부 평균(conditional mean)이 된다.

$$E_Y[Y|X] = \int_{y=-\infty}^{y=\infty} y\, p_{Y|X}(y|x)\, dy$$

또는 간단히 다음처럼 쓴다.

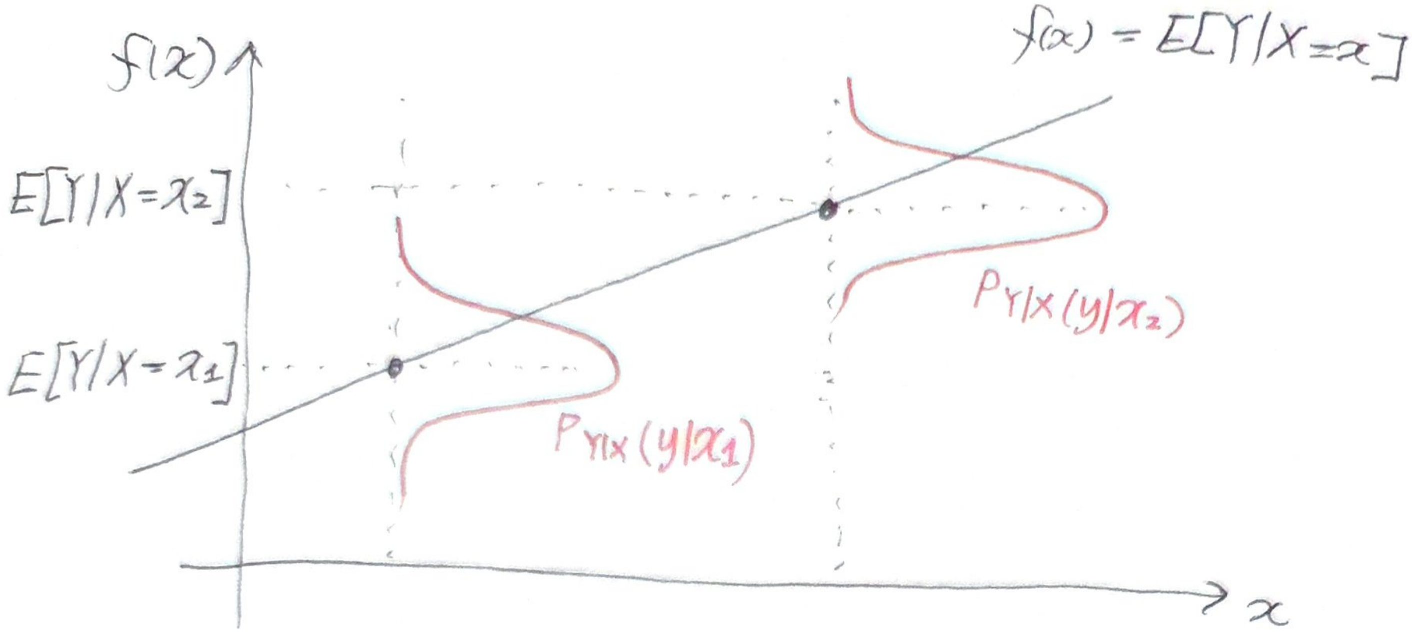
$$E[Y|X] = \int y\, p(y|x)\, dy$$

또는 간단히 다음처럼 쓴다.

$$E[Y|X] = \int y\, p(y|x)\, dy$$

조건부 기댓값에서 조건이 되는 확률변수 X의 값 x는 조건부 기댓값을 사용하는 사용자가 지정해야 하는 독립변수이다.
따라서 조건부 기댓값은 조건이 되는 확률변수의 값에 따라서 값이 달라지는 확률변수이다. 또한 $E[Y|X]$는 조건이 되는 확률변수 X의 값 x를 입력으로 갖는 함수다.

$$E[Y|\overset{\text{user input}}{\boxed{X=x}}] = f(x)$$



x에 따라 결정되는 $f(x) = E[Y|X=x]$

= x에 따라 결정되는 확률분포의 중심값, 중심위치.

= $\boxed{\text{조건부 기댓값}}$

* $E[Y]$ = 결정론적, 상수 = $\int y P(y)\, dx$

* $E[Y|X]$ = x에 따라 결정 $\boxed{\text{확률변수}}$ ∨ = $\int y P(y|x)\, dx$

x로 결정되는 숫자

$f(\underline{\underline{x}})$ ... x도 확률변수 (X의 변환)

확률변수 X의 출력값

$f(X)$ ... "확률변수의 변환"

예측문제

두 확률변수 X, Y에서 X값을 알고 있을 때 Y값을 추정하기

= $\boxed{\text{예측 (prediction)}}$.

- Y가 연속확률변수 ⟶ regression analysis
- Y가 이산확률변수 ⟶ classification

일반적으로 조건부기댓값 = 예측문제의 답 = $\hat{y}$
(때에 따라 최빈값, 중앙값)

$$x \xrightarrow{\text{prediction}} \hat{y} = E[y|x] = f(x)$$

조건부기댓값의 성질

조건부기댓값 $E[Y|X]$가 X의 함수 (=변환, transformation)이므로,
조건부기댓값 $E[Y|X]$도 확률변수다.

if 확률변수 Y가 확률변수 X값을 독립변수로 하는 결정론적 함수값이라면

$$Y = g(X)$$

사용자가 X의 값을 어떤 x로 정하면 Y가 결정되므로 $Y=g(x)$는 상수가 된다.

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

같은 방식으로, 확률변수 X와 Y가 결정론적 함수관계가 아닐 때도 다음 등식이 성립한다.

$$E[\underbrace{g(X)}_{\text{상수}}Y|X] = g(X)E[Y|X]$$

전체 기댓값의 성질

조건부 기댓값은 확률변수이므로 조건이 되는 확률변수에 대해 다시 기댓값을 구할 수 있다.
이렇게 반복하여 구한 조건부 기댓값의 기댓값은 원래 확률변수의 기댓값과 같다.

$$E_X[\underbrace{E_Y[Y|X]}_{\text{확률변수}}] = E_Y[Y]$$

간단히 다음처럼 쓰기도 한다.

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

이를 **전체 기댓값의 법칙(law of total expectation)** 또는
반복 기댓값의 법칙 (law of iterated expectation) 이라고 한다.

X, Y가 이산확률변수인 경우에는 다음처럼 증명할 수 있다.

$$E_X[E_Y[X|Y]] = \sum_{x_i \in X} p(x_i) E_Y[X|Y]$$

$$= \sum_{x_i \in X} p(x_i) \sum_{y_i \in Y} p(y_i | x_i) y_i$$

$$= \sum_{x_i \in X} \sum_{y_i \in Y} p(x_i) p(y_i | x_i) y_i = \sum_{x_i \in X} \sum_{y_i \in Y} p(x_i, y_i) y_i$$

전체확률의 법칙.

$$= \sum_{y_i \in Y} p(y_i) y_i$$

$$= E_Y[Y]$$

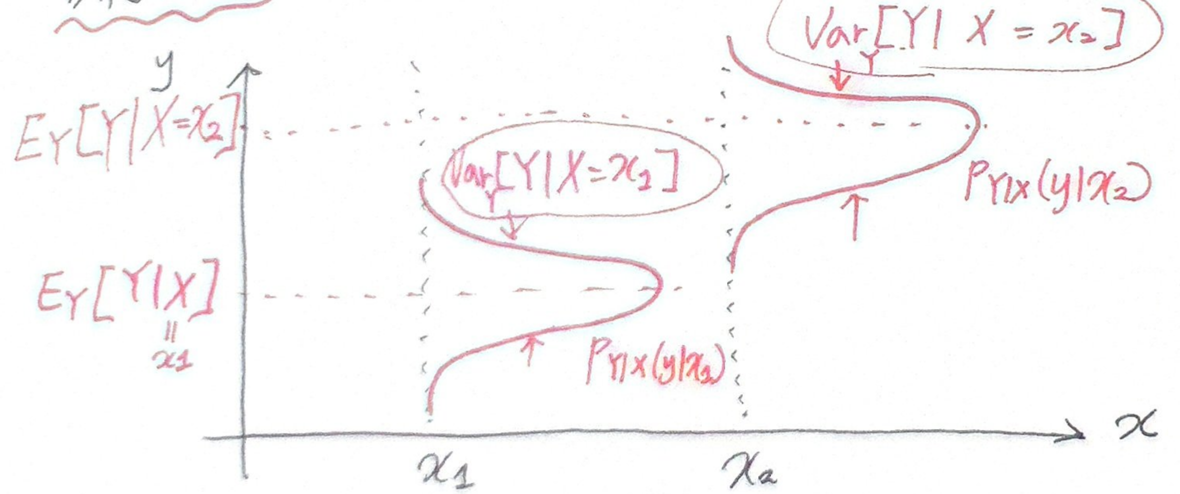
조건부분산

조건부기댓값을 정의한 것처럼 조건부분산(conditional variance)도 다음처럼 정의할 수 있다.

$$Var_Y[Y|X] = E_Y[(Y-E_Y[Y|X])^2|X]$$

$$= \int (Y-E_Y[Y|X])^2 f_{Y|X}(y|x)\,dy$$

조건부분산은 x 값을 알고 있을 때 이에 대한 조건부확률분포 $p(y|x)$의 분산이다. 예측문제의 관점으로 보면 조건부분산은 예측의 불확실성, 즉 예측으로 맞출수 없는 범위를 뜻한다.



전체 분산의 법칙

확률변수의 분산은 조건부분산의 기댓값과 조건부기댓값의 분산의 합과 같다. 이를 전체 분산의 법칙(law of total variance)이라고 한다.

$$\boxed{\underbrace{Var}_{\text{분산}}[Y]} = \underbrace{\underbrace{E}_{\text{기댓값}}[\underbrace{Var[Y|X]}_{\text{조건부분산}}]}_{\text{회귀분석에서 나올수 있는 예측오차의 평균}} + \underbrace{\underbrace{Var}_{\text{분산}}[\underbrace{E[Y|X]}_{\text{조건부기댓값}}]}_{\text{조건부기댓값의 변동폭}}$$

① 회귀분석에서 나올수있는 예측오차 평균

② 회귀분석 예측치의 변동폭

(증명)

- 전체 기댓값의 법칙을 사용.

$$
\begin{aligned}
Var[Y] &= E[Y^2] - (E[Y])^2 \\
&= E[E[Y^2 | X]] - (E[E[Y|X]])^2 \\
&= E[Var[Y|X] + (E[Y|X])^2] - (E[E[Y|X]])^2 \\
&= E[Var[Y|X]] + (E[(E[Y|X])^2] - (E[E[Y|X]])^2) \\
&= E[Var[Y|X]] + Var[E[Y|X]]
\end{aligned}
$$

$E[Y|X] = \hat{y}$로 표현하면,

$$Var[Y] = E[(\hat{y} - y)^2] + Var[\hat{y}]$$

☆ Takeaways

- 예측문제의 관점에서, 조건부 분산의 기댓값 $E[(\hat{y}-y)^2]$은 예측 오차, 즉 편향(bias)의 평균적인 크기를 뜻한다.
- 조건부 기댓값의 분산 $Var[\hat{y}]$은 예측값의 변동 크기다.
- 예측값의 변동 크기가 증가한다 == 예측모형이 복잡하고 비선형적이며, 주어진 데이터에 과적화되기 쉽다

따라서,

(1) 예측 오차의 크기와 예측값 변동 합이 일정하므로, 예측 오차를 줄이면 모형이 복잡해지고 과최적화가 되며

(2) 반대로 모형 과최적화를 막기 위해 단순하게 하면 예측 오차 증가

⇒ 편향-분산 상충 (Bias-variance Tradeoff)