

Lecture 13: Conditional Expectation & Information Theory**Conditional Expectation**

$$\begin{aligned}
E[Y|X = x] &= \text{expectation of } Y \text{ given } X = x \\
&= \sum_{j=1}^J y_j P_{Y|X=x}(y_j) \\
&= \sum_{j=1}^J y_j \frac{P(X = x, Y = y_j)}{P(X = x)}
\end{aligned}$$

Let $g(x) = E[Y|X = x]$ then $g(X) = E[Y|X]$ is a function of r.v. X , so a r.v. itself.

Fact: $E[Y] = E\{E(Y|X)\}$

Proof:

$$\begin{aligned}
\text{RHS} &= \sum_{i=1}^I P(X = x_i) E[Y|X = x_i] \\
&= \sum_{i=1}^I P(X = x_i) \sum_{j=1}^J y_j \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} \\
&= \sum_{j=1}^J y_j \sum_{i=1}^I P(X = x_i, Y = y_j) \\
&= \sum_{j=1}^J y_j P(Y = y_j) = EY = \text{LHS}
\end{aligned}$$

Best prediction of Y based on X in terms of MSE

Find $\Phi(X)$ that minimizes $E[Y - \Phi(X)]^2$.

$$E[Y - \Phi(X)]^2 = \sum_{i=1}^I P(X = x_i) E \left[[Y - \Phi(x_i)]^2 | X = x_i \right]$$

The i^{th} term in the summation is minimized by $E[Y|X = x_i]$ so best $\Phi(X) = E[Y|X]$.

Best linear predictor in first Hexamer example:

$$EY + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - EX) = 0.570 + 0.164(X - 0.505)$$

For $X = 1$, best linear predictor for Y is 0.651. $E[Y|X = 1] = 0.46$ — best predictor.

If X_i iid $N(\mu_X, \sigma_X^2)$ and Y_i follow the simple linear regression model, i.e.,

$$Y_i = a_0 + b_0 X_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \text{ iid and independent of } \{X_i\}$$

$$E[Y|X] = a_0 + b_0 X$$

then (X_i, Y_i) bivariate Normal

1. $\mu_Y = a_0 + b_0 \mu_X$
2. $\text{Var}(Y) = b_0^2 \sigma_X^2 + \sigma^2$
3. $\text{Cov}(X, Y) = b_0 \sigma_X^2$

$$4. \rho_{XY} = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)} = \frac{b_0 \sigma_X^2}{\sigma_X \sqrt{b_0^2 \sigma_X^2 + \sigma^2}} = \frac{b_0 \sigma_X}{\sqrt{b_0^2 \sigma_X^2 + \sigma^2}}$$

5. Best linear predictor, $a + bX$

$$6. |\rho_{XY}| \leq 1, \quad |\rho_{XY}| \equiv 1 \iff \sigma = 0$$

Fact: If X_1, \dots, X_n independent Normal with distribution $N(\mu_i, \sigma^2)$, $i = 1, \dots, n$ then,

$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma^2\right)$$

Proof:

$$1. Y = a_0 + b_0 X + \epsilon, \quad EY = a_0 + b_0 EX + E\epsilon = a_0 + b_0 EX$$

$$2. \text{Var}(Y) = \text{Var}(a_0 + b_0 X + \epsilon) = \text{Var}(b_0 X + \epsilon)$$

$$= b_0^2 \text{Var}(X) + \text{Var}(\epsilon)$$

$$= b_0^2 \sigma_X^2 + \sigma^2$$

$$3. \text{Cov}(X, Y) = \text{Cov}(X, a_0 + b_0 X + \epsilon)$$

$$= \text{Cov}(X, a_0) + \text{Cov}(X, b_0 X) + \text{Cov}(X, \epsilon)$$

$$= 0 + b_0 \text{Cov}(X, X) + 0 = b_0 \sigma_X^2$$

4. Plug-in

$$5. \text{Best linear predictor} = E[Y|X] = a + bX = a_0 + b_0 X$$

$$a = EY - bEX$$

$$b = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = b_0$$

$$6. |\rho_{XY}| \equiv 1 \iff b_0^2 \sigma_X^2 = b_0^2 \sigma_X^2 + \sigma^2 \iff \sigma^2 = 0$$

Elements of Information Theory

Ref: Cover and Thomas, 1991

Another way to measure randomness or variability in a r.v. X is through Shannon's Entropy (values of X do not matter)

$$H(X) = -\sum_{i=1}^m p_i \log_2 p_i = -E \log_2 p(X) \quad p_i = P(X = x_i), \quad i = 1, \dots, m$$

the operational meaning of $H(X)$ is given by Shannon's Source Coding Theorem.

If X_1, \dots, X_n iid, then for any binary prefix code $C : (X_1, \dots, X_n) \rightarrow \{0, 1\}^*$

$$\frac{EL_n(X_1, \dots, X_n)}{n} \geq H(X)$$

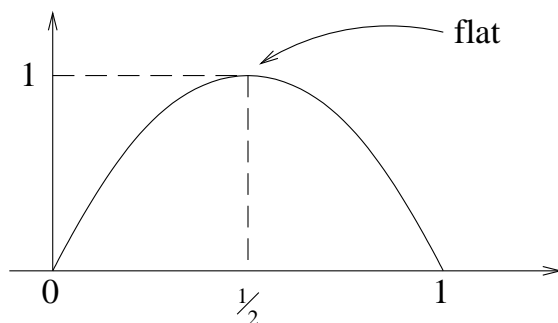
$L_n(X_1, \dots, X_n)$ is the bits in the codeword of (X_1, \dots, X_n)

Moreover, the Shannon code with codelength $L_n(X_1, \dots, X_n) = \lceil -\log_2 p(X_1, \dots, X_n) \rceil$ achieves the entropy lower bound.

Examples:

1. X_1, \dots, X_n iid Bernoulli(p)

$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p) \quad (\text{bits})$$



$p = 1/2$	$H(X) = 1$ bits
$p = 1/3$	$H(X) = 0.92$ bits
$p = 1/4$	$H(X) = 0.81$ bits
$p = 1/10$	$H(X) = 0.47$ bits

2. $X_1 \dots X_n$ iid uniform on $\{x_1, \dots, x_m\}$

$$H(X) = \log_2 m$$

3. Differential entropy for continuous r.v. X .

$$H(X) = - \int_{-\infty}^{\infty} f(x) \log_2 f(x) dx$$

If $X \sim N(\mu, \sigma^2)$, $H(X) = \frac{1}{2} \log_2(2\pi e \sigma^2)$ bits

Joint Entropy and Conditional Entropy: (X, Y)

- **Joint Entropy** $H(X, Y)$ is defined as:

$$H(X, Y) = - \sum_{x_i} \sum_{y_j} p_{ij} \log_2 p_{ij} = -E \log P(X, Y)$$

First Hexamer example:

$$\begin{aligned} H(X, Y) &= -(0.424 \log_2 0.424 + 0.180 \log_2 0.180 + \dots + 0.002 \log_2 0.002) \\ &= 2.79 \text{ bits} \end{aligned}$$

(NB: $0 \log_2 0$ is taken to equal 0)

Second Hexamer example:

$$\begin{aligned} H(X, Y) &= -(0.158 \log_2 0.158 + 0.213 \log_2 0.213 + \dots + 0.002 \log_2 0.002) \\ &= 3.56 \text{ bits} \end{aligned}$$

3.56 bits > 2.79 bits — surprise? No ...

- **Conditional Entropy:**

$$\begin{aligned} H(Y|X) &= -E_{p(x,y)} \log_2 P(Y|X) \\ &= - \sum_{x_i} \sum_{y_j} p(x_i, y_j) \log_2 P(y_j|x_i) \end{aligned}$$

- **Chain Rule:**

$$H(X, Y) = H(X) + H(Y|X)$$

If X, Y are independent, $H(Y|X) = H(Y)$

$$H(X, Y) = H(X) + H(Y)$$

- **Entropy rate of a stationary process:**

$$H(X^\infty) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1 \dots X_n)$$

where $X^\infty = (X_1, \dots, X_n, \dots)$. $H(X^\infty)$ gives the compression limit of a stationary process, similar as in the iid case.

Relative Entropy & Mutual Information

- A measure of “distance” between two distributions: it’s also called the Kullback-Leibler (KL) divergence between two distributions $p(x)$ and $q(x)$:

$$\begin{aligned} D(p||q) &= \sum_i p_i \log \frac{p_i}{q_i} \\ &= E_p \log \frac{p(X)}{q(X)} \geq 0 \\ D(q||p) &\neq D(p||q), \quad D(p||p) = 0 \end{aligned}$$

- Mutual Information $I(X; Y)$

$$I(X; Y) = D(p(x, y)||p(x)q(y))$$

$$I(X; Y) = 0 \iff X, Y \text{ are independent}$$

It measures how much information X and Y contain about each other.

In the first hexamer example: $H(X) = 1.37$, $H(Y) = 1.46$, $H(X, Y) = 2.79$

$$I(X; Y) = 0.037 \text{ bits}$$

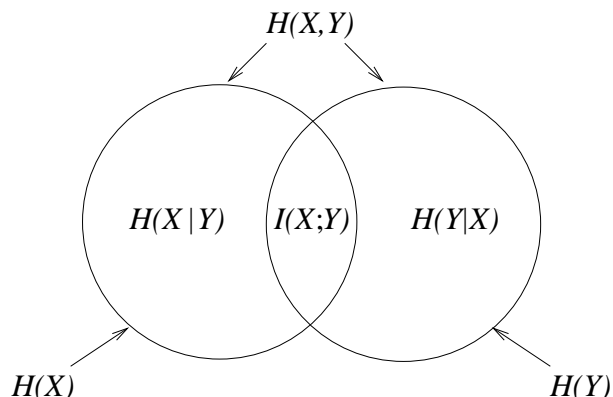
In the second hexamer example: $H(X) = 1.41$, $H(Y) = 2.16$, $H(X, Y) = 3.56$

$$I(X; Y) = 0.012 \text{ bits}$$

- Relationship between Mutual Information and Entropy

$$I(X; Y) = H(X) - H(X|Y)$$

$$I(X; Y) = H(Y) - H(Y|X)$$



$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$