

1. Joint, Marginal, and Conditional Probabilities

1.1 Compute $P(X \leq 2, Y > 1)$.

$$P(X \leq 2, Y > 1) = P(X = 1, Y = 2) + P(X = 2, Y = 2) = \frac{1}{12}$$

1.2 Compute marginal probability mass function for X and Y .

IDKMANDOTHS

1.3 Compute $P(Y = 2|X = 1)$.

$$P(Y = 2|X = 1) = \frac{P(Y = 2, X = 1)}{P(X = 1)} = \frac{1/12}{1/3 + 1/12} = \frac{1}{5}$$

1.4 Are X and Y independent?

$$P(X = 1) = \frac{5}{12} \neq \frac{4}{7} = P(X = 1|Y = 1)$$

Thus they are not independent.

1.5 Define $Z = X - 2Y$, compute $P(X = 2|Z = 0)$.

$$P(X = 2|Z = 0) = \frac{P(X = 2, Z = 0)}{P(Z = 0)}$$

$P(Z = 0) = P(X = 2, Y = 1) + P(X = 4, Y = 2) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$ and $P(X = 2, Z = 0) = P(X = 2, Y = 1) = \frac{1}{6}$ so

$$P(X = 2|Z = 0) = \frac{1/6}{1/2} = \frac{1}{3}$$

1.6 Compute $E[X|Y = 1]$. First compute $P(Y = 1) = \frac{1}{3} + \frac{1}{6} + \frac{1}{12} = \frac{7}{12}$.

$$\begin{aligned} E[X|Y = 1] &= 1(P(X = 1|Y = 1) + 2(P(X = 2|Y = 1)) + 4(P(X = 4|Y = 1))) \\ &= 1\left(\frac{P(X = 1, Y = 1)}{P(Y = 1)}\right) + 2\left(\frac{P(X = 2, Y = 1)}{P(Y = 1)}\right) + 4\left(\frac{P(X = 4, Y = 1)}{P(Y = 1)}\right) \\ &= \frac{1/3}{7/12} + 2\left(\frac{1/6}{7/12}\right) + 4\left(\frac{1/12}{7/12}\right) \\ &= \frac{12}{7} \end{aligned}$$

1.7 Compute $\text{Var}[X|Y = 2]$. $\text{Var}[X|Y = 2] = E[X^2|Y = 2] - E[X|Y = 2]^2$ First compute $P(Y = 2) = \frac{1}{12} + \frac{1}{3} = \frac{5}{12}$.

$$\begin{aligned} E[X|Y = 2] &= 1(P(X = 1|Y = 2) + 2(P(X = 2|Y = 2)) + 4(P(X = 4|Y = 2))) \\ &= 1\left(\frac{P(X = 1, Y = 2)}{P(Y = 2)}\right) + 2\left(\frac{P(X = 2, Y = 2)}{P(Y = 2)}\right) + 4\left(\frac{P(X = 4, Y = 2)}{P(Y = 2)}\right) \\ &= \frac{1/12}{5/12} + 4\left(\frac{1/3}{5/12}\right) \\ &= \frac{17}{5} \end{aligned}$$

and

$$\begin{aligned} E[X^2|Y = 2] &= 1(P(X = 1|Y = 2) + 4(P(X = 2|Y = 2)) + 16(P(X = 4|Y = 2))) \\ &= 1\left(\frac{P(X = 1, Y = 2)}{P(Y = 2)}\right) + 4\left(\frac{P(X = 2, Y = 2)}{P(Y = 2)}\right) + 16\left(\frac{P(X = 4, Y = 2)}{P(Y = 2)}\right) \\ &= \frac{1/12}{5/12} + 16\left(\frac{1/3}{5/12}\right) \\ &= 13 \end{aligned}$$

$$\text{so } \text{Var}[X|Y = 2] = 13 - \frac{17^2}{5} = 1.44$$

Solution

2. Proof of Probabilistic Ranking Principle

Solution

Assume that the system returns n results. If the user reads all of the results or none of the results then it is clear the ranking doesn't matter. So we assume that the user reads some number $r \in \mathbb{N}$ of the results and then stops. We will show that changing the ordering of the results (by swapping two results) can only make the ordering worse from a risk minimization perspective. First consider the risk of presenting an irrelevant document. This risk is calculated as:

$$\prod_{i=0}^r 1 - p(R = 1|Q, D_i)$$

where D_1 is the first result presented, D_2 the second, and so on. Consider what happens when we swap the ordering of two documents D_j and D_k . If $j, k \leq r$ or if $j, k > r$ then the risk doesn't change. So we only care what happens when one is less than or equal to r and the other is greater. Without loss of generality assume $j \leq r$ and $k > r$. In calculating the risk we've replaced the term $1 - p(R = 1|Q, D_j)$ with $1 - p(R = 1|Q, D_k)$. By our ordering we know that $1 - p(R = 1|Q, D_j) \leq 1 - p(R = 1|Q, D_k)$. Thus the risk either increases or stays the same as a result of the swap.

Now we look at the risk of missing a relevant document. This risk is computed as:

$$\prod_{i=r+1}^n p(R = 1|Q, D_i)$$

Consider what happens when we swap the ordering of two documents D_j and D_k . If $j, k \leq r$ or if $j, k > r$ then the risk doesn't change. So we only care what happens when one is less than or equal to r and the other is greater. Without loss of generality assume $j \leq r$ and $k > r$. In calculating the risk we've replaced the term $p(R = 1|Q, D_j)$ with $p(R = 1|Q, D_k)$. By our ordering we know that $p(R = 1|Q, D_j) \geq p(R = 1|Q, D_k)$. Thus the risk either increases or stays the same as a result of the swap.

3. Maximum Likelihood Estimation

Solution

The likelihood function is:

$$\begin{aligned} L(\mu, \sigma) &= f(x_1|\mu, \sigma) \times \cdots \times f(x_n|\mu, \sigma) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \times \cdots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2} \end{aligned}$$

Take the log to get:

$$L(\mu, \sigma) = -\frac{n}{2}(\ln 2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

To maximize this function we take the derivative and then set it equal to zero. First we take the derivative with respect to μ .

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = \frac{1}{\sigma^2} \sum_{i=1}^n x_i - n\mu$$

.

In order for this to be zero we must have $\hat{\mu} = \bar{x}$.

We repeat the process for σ . We take the derivative by σ^2 to simplify computations.

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{(2\sigma^2)^2} \left[\sum_{i=1}^n (x_i - \mu)^2 \right] = \frac{1}{2\sigma^2} \left[\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - n \right]$$

.

In order for this to be zero we must have

$$\begin{aligned}\frac{1}{2\sigma^2} \left[\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] &= n \\ \implies \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \\ \implies \hat{\sigma} &= \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2}\end{aligned}$$

4. Cosine vs. Euclidean Distance

Solution

Let $x = \{x_1, \dots, x_n\}$ and $y = \{y_1, \dots, y_n\}$. Then

$$C(x, y) = \frac{x \cdot y}{|x||y|}$$

We are dealing with unit vectors so the denominator is zero and

$$C(x, y) = x \cdot y = \sum_{i=1}^n x_i y_i$$

Euclidean distance is given by $E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$. Then to discover the relationship between C and E consider $E(x, y)^2$.

$$\begin{aligned}E(x, y)^2 &= \sum_{i=1}^n (x_i - y_i)^2 \\ &= \sum_{i=1}^n x_i^2 + y_i^2 - 2x_i y_i \\ &= \sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2 - 2 \sum_{i=1}^n x_i y_i \\ &= 2 - 2 \sum_{i=1}^n x_i y_i \\ &= 2(1 - C(x, y))\end{aligned}$$

5. α_d in Language Model Ranking Modules

Solution

By definition if w is not seen in d then $p(w|d) = \alpha_d p(w|REF)$. For Dirichlet prior smoothing

$$p(w|d) = \frac{|d|}{|d| + \mu} \frac{c(w, d)}{|d|} + \frac{\mu}{|d| + \mu} p(w|REF)$$

In the case of an unseen word $c(w, d) = 0$. This implies that

$$p(w|d) = \alpha_d p(w|REF) = \frac{\mu}{|d| + \mu} p(w|REF) \implies \alpha_d = \frac{\mu}{|d| + \mu}$$