2.6. Probability and Statistics

- Probability: Reasoning under uncertainty (given a probabilistic model of a process, we can reason about the likelihood of different events)

- Statistics: the study of data: collecting, analyzing, interpreting, and drawing conclusions from datasets. It often involves making unknown patterns of a population based on a sample.

2.6.1. Example: Tossing coins

- Supposed the coin is fair (P(head) = 0.5), we can simulate multiple draws with the Multinomial function
- Each time you run this sampling process, you get a different result
- As the number of samples grows, the sample estimates converge to the true underlying probabilities (Central Limit Theorem).

2.6.2. Formal notations

- Set of possible outcomes (sample space) $S = \{heads, tails\}$ if the task is tossing a coin.
- If we're tossing 2 coins: $S = \{(heads, heads), (heads, tails), (tails, heads), (tails, tails)\}$

+ Example: rolling a dice: $S = \{1, 2, 3, 4, 5, 6\}$

- Given a random variable $X, P(X = v)$ denotes the probability of $X\ taking\ value\ v$
- Similarly, $P(1 \leq X \leq 3)$ indicates the probability of event $\{1 \leq X \leq 3\}$

- A probability function P maps events onto real values:

$$P: A \subseteq S \ \rightarrow [0,1]$$

- The probability, denoted $P(A)$, of an event A in sample space S has the following properties:

    1. The probability of any event A is a real non-negative number:

    $$P(A) \geq 0$$

    2. The probability of the entire sample space is 1:

    $$P(S) = 1$$

    3. For any sequence of events $A_1, A_2, \ldots$ that are mutually exclusive ($A_i \cap A_j = \emptyset \ \ for\ all\ i \neq j$), the probability that any of them happen is equal to the sum of their individual probabilities:

    $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

2.6.3. Random variables

- 2 types: discreet and continuous
- Example:

+ X is the number rolled on a dice (discreet)

+ Y is the height of a group sampled at random from a population (continuous)

Let X be the exact amount of rain tomorrow:

$$P(X = 2) = ?$$

Probability density function $p(x)$ with $P(X) = \int_{-\infty}^{\infty} p(x)dx$

Example: $P(X \leq 2) = \int_{0}^{2} p(x)dx$

2.6.4. Multiple random variables

- **Joint probability** $P(A = a, B = b)$ denotes the probability of event $A = a\ and\ B = b$ happening at the same time:

$$P(A = a, B = b) \leq P(A = a)$$
$$P(A = a, B = b) \leq P(B = b)$$

+ To get P(A=a), take sum of all $P(A = a, B = v)$ with all values v that random variable B can get :

$$P(A = a) = \sum_v P(A = a, B = v)$$

- **Conditional probability** $P(A = a|B = b)$ denotes the probability of event $A = a$, once the condition $B = b$ is met

$$P(A = a, B = b) = \frac{P(A = a, B = b)}{P(B = b)}$$

+ For 2 disjoint events B and B' : $P(B \cup B'|A = a) = P(B|A = a) + P(B'|A = a)$

Bayes theorem

- With the conditional probability equation, we have:

$$P(A, B) \; = \; P(A|B)P(B) \; = \; P(B|A)P(A)$$

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

P(A|B): posterior

P(B|A): likelihood

P(A): prior

P(B): evidence

- Example: if we know the prevalence of symptoms for a disease, we can determine how likely someone has the disease based on the symptoms.
- In case we don't have access to $P(B)$, a simpler version of Bayes theorem can be used:

$$P(A|B) \propto P(B|A)P(A)$$

- Since $P(A|B)$ must be normalized to 1, meaning $\sum_a P(A = a|B) = 1$, we also have:

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

$\sum_a P(B|A = a)P(A = a)$ = $\sum_a P(B|A = a) = P(B)$

Independence

- Random variables A and B are independent if changes on value of A does not change the probability distribution of B and vice versa.

A, B are independent $(A \perp B)$

$\rightarrow P(A|B) = P(A) \rightarrow P(A,B) = P(A|B)P(B) = P(A)P(B)$

- **Conditional Independence**: random variables A and B are conditionally independent given a third variable C iff $P(A,B|C) = P(A|C)P(B|C)$
- Example: broken bones and cancer are independent if we consider the whole population. However, if we condition on being in a hospital, broken bones are negatively correlated with having cancer.

Example: Doctor administer HIV test to a patient. $D_1 = 1$ means positive and $D_1 = 0$ means negative. H is the HIV status of the patient. Assume P(H=1) = 0.0015

$$P(H = 1|D1 = 1) = ?$$

$$\begin{aligned} P(D1 = 1) &= P(D1 = 1, H = 0) + P(D1 = 1, H = 1) \\ &= P(D1 = 1|H = 0)\, P(H = 0) + P(D1 = 1|\, H = 1)\, P(H = 1) \\ &= 0.01 \times 0.9975 + 1 \times 0.0015 \\ &= 0.011475 \end{aligned}$$

Using Bayes rules:

$\rightarrow P(H = 1|D1 = 1) = \frac{P(D1=1|H=1)P(H=1)}{P(D1=1)} = \frac{0.0015}{0.0011475}$ = 0.1306

→ There's 13% chance the patient have HIV if diagnosed positive, even though the test is very accurate according to the table.

This is counter-intuitive

Conditional probability	$H = 1$	$H = 0$
$P(D_1 = 1 \mid H)$	1	0.01
$P(D_1 = 0 \mid H)$	0	0.99

- Second test is not as accurate as the first one

$$P(D2\ = 1)\ =\ 0.98 \times 0.0015\ +\ 0.03 \times\ 0.9975\ =\ 0.0314$$

$$P(H = 1|D2 = 1)\ = \frac{0.98 \times 0.0015}{0.0314}\ =\ 0.0468$$

Conditional probability	$H = 1$	$H = 0$
$P(D_2 = 1 \mid H)$	0.98	0.03
$P(D_2 = 0 \mid H)$	0.02	0.97

Second test also came out positive with 4.68% of getting HIV.

Assuming conditional independence for test 1 and 2, we have:

$$P(D1 = 1, D2 = 1|H = 0) = P(D1 = 1|H = 0)\ P(D2 = 1|H = 0) = 0.0003$$

$$P(D1 = 1, D2 = 1|H = 1) = P(D1 = 1|H = 1)\ P(D2 = 1|H = 1) = 0.98$$

$$\begin{aligned}
&P(D1 = 1, D2 = 1) \\
&= P(D1 = 1, D2 = 1, H = 0) + P(D1 = 1, D2 = 1, H = 1) \\
&= P(D1 = 1, D2 = 1|H = 0)\ P(H = 0) + P(D1 = 1, D2 = 1|H = 1)\ P(H = 1) \\
&= 0.00177
\end{aligned}$$

$$P(H = 1|D1 = 1, D2 = 1) = \frac{P(D1 = 1, D2 = 1|H = 1)\ P(H = 1)}{P(D1 = 1, D2 = 1)} = 0.8307$$

The second test significantly improved the estimate when combined with the first one

2.6.6 Expectations

- Expectation of random variable X is defined as:

$$E[X] = E_{x \sim P}[x] = \sum_x xP(X = x)$$

- For densities, we have $E[X] = \int x \, dp(x)$
- Expected value of some function f(x):

$$E_{x \sim P}[f(x)] = \sum_x f(x)P(x) = \int f(x)p(x)dx$$

Variance

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

- The variance of a function of a random variable:

$$Var_{x\sim P}[f(x)] = E_{x\sim P}[f^2(x)] - E_{x\sim P}f(x)^2$$

- Standard deviation:

$$\sigma = \sqrt{Var(X)}$$

Expectation and variance of vector

Apply the formula elementwise:

$$\boldsymbol{\mu} \stackrel{\text{def}}{=} E_{\boldsymbol{x}\sim P}[\boldsymbol{x}]$$

$\boldsymbol{\mu}$ has coordinates $\mu_i = E_{\mathbf{x}\sim P}[x_i]$

Covariance matrix:

$$\boldsymbol{\Sigma} \stackrel{\text{def}}{=} Cov_{\mathbf{x}\sim P}[\mathbf{x}] = E_{\mathbf{x}\sim P}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$$

Let $\mathbf{v}$ be a vector of the same size as $\mathbf{x}$

$$\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v} = E_{\mathbf{x}\sim P}[\mathbf{v}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{v}] = Var_{\mathbf{x}\sim P}[\mathbf{v}^{\mathrm{T}}\mathbf{x}]$$

$\boldsymbol{\Sigma}$ allows us to compute variance for any linear function of $\mathbf{x}$ with matrix multiplication. The off-diagonal elements show the correlation between coordinates.

0 means low correlation, large positive value means they are strongly correlated

Maximum likelihood

- Suppose we have a model with parameters $\boldsymbol{\theta}$ and data samples X, we want to find the most likely value for the parameters:

$$\operatorname{argmax} P(\boldsymbol{\theta} \mid X).$$

Using Bayes rules:

$$\operatorname{argmax} \frac{P(X \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(X)}.$$

$P(X), P(\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$ (uninformative prior)

$\rightarrow$ $$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} P(X \mid \boldsymbol{\theta}).$$

The probability of the data given the parameter $P(X|\theta)$ is called the likelihood

Numerical Optimization and Negative log-likelihood

- Instead of finding $argmax_\theta P(X|\boldsymbol{\theta})$ we can find $argmax_\theta \log(P(X|\boldsymbol{\theta}))$, since log(x) is a monotone increasing function

$argmax_\theta \log(P(X|\boldsymbol{\theta}))$ = $argmin_{\boldsymbol{\theta}} - \log(P(X|\boldsymbol{\theta}))$

- Related to information theory, entropy is the amount of randomness in a random variable

$$H(p) = -\sum_i p_i \log_2(p_i),$$

- If we take the negative log-likelihood and divide by n samples, we get cross-entropy (a way to measure classification performance)

- Due to independence assumption, most probabilities we see in ML are products of individual probabilities:

$$P(X \mid \boldsymbol{\theta}) = p(x_1 \mid \boldsymbol{\theta}) \cdot p(x_2 \mid \boldsymbol{\theta}) \cdots p(x_n \mid \boldsymbol{\theta}).$$

- Using the product rule to compute derivative

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} P(X \mid \boldsymbol{\theta}) = & \left(\frac{\partial}{\partial \boldsymbol{\theta}} P(x_1 \mid \boldsymbol{\theta})\right) \cdot P(x_2 \mid \boldsymbol{\theta}) \cdots P(x_n \mid \boldsymbol{\theta}) \\
& + P(x_1 \mid \boldsymbol{\theta}) \cdot \left(\frac{\partial}{\partial \boldsymbol{\theta}} P(x_2 \mid \boldsymbol{\theta})\right) \cdots P(x_n \mid \boldsymbol{\theta}) \\
& \vdots \\
& + P(x_1 \mid \boldsymbol{\theta}) \cdot P(x_2 \mid \boldsymbol{\theta}) \cdots \left(\frac{\partial}{\partial \boldsymbol{\theta}} P(x_n \mid \boldsymbol{\theta})\right).
\end{aligned}$$

- This needs n(n-1) multiplications, so it’s proportional to quadratic time in the inputs (inefficient).

- Instead we can use negative log-likelihood

$$-\log\left(P(X \mid \boldsymbol{\theta})\right) = -\log(P(x_1 \mid \boldsymbol{\theta})) - \log(P(x_2 \mid \boldsymbol{\theta})) \cdots - \log(P(x_n \mid \boldsymbol{\theta})),$$

- Compute derivative:

$$-\frac{\partial}{\partial \boldsymbol{\theta}} \log\left(P(X \mid \boldsymbol{\theta})\right) = \frac{1}{P(x_1 \mid \boldsymbol{\theta})}\left(\frac{\partial}{\partial \boldsymbol{\theta}} P(x_1 \mid \boldsymbol{\theta})\right) + \cdots + \frac{1}{P(x_n \mid \boldsymbol{\theta})}\left(\frac{\partial}{\partial \boldsymbol{\theta}} P(x_n \mid \boldsymbol{\theta})\right).$$

- This needs n divisions and n sums -> Linear time

Example

- Given $X = \{x_i\}_{i=1}^{n}$ is a random sample from an exponential distribution with parameter $\lambda > 0$. It has the following p.d.f:

$$p(x) = \lambda e^{-\lambda x}$$

The likelihood is: $L(X|\lambda) = \prod_{i=1}^{n} p(x_i|\lambda)$

$$= \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

$$= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

We want to find the maximum likelihood estimate:

$$\widehat{\lambda}_n(\boldsymbol{x}) = \arg\max_{\lambda \in \mathbb{R}^+} \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\widehat{\lambda}_n(\boldsymbol{x}) = \arg\max_{\lambda \in \mathbb{R}^+} \left\{ n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i \right\}$$

We can find the maximum by taking the derivative and equate to 0:

$$\frac{\partial}{\partial \lambda} \left(n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i \right) = 0 \iff \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \ .$$

Also $\frac{\partial^2}{\partial \lambda^2} \left(n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i \right) = -\frac{n}{\lambda^2} < 0$, so the solution of the above equation is indeed the global maximum.

The maximum likelihood estimate is $$\widehat{\lambda}_n(\boldsymbol{x}) = \frac{n}{\sum_{i=1}^{n} x_i} \ .$$

- Given X = [2.7, 4.9, 0.2, 4.9, 4.4, 18.7, 1.5, 0.9, 10.5, 1.3] following an exponential distribution (n = 10)

-> The MLE is $$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{10}{50} = \frac{1}{5} = 0.2 \ .$$

Maximum likelihood for Continuous variables

- For continuous variables we want to compute within a range ϵ

$$
\begin{aligned}
&P(X_1 \in [x_1, x_1+\epsilon], X_2 \in [x_2, x_2+\epsilon], \ldots, X_N \in [x_N, x_N+\epsilon] \mid \boldsymbol{\theta}) \\
\approx &\epsilon^N p(x_1 \mid \boldsymbol{\theta}) \cdot p(x_2 \mid \boldsymbol{\theta}) \cdots p(x_n \mid \boldsymbol{\theta}).
\end{aligned}
$$

- Take negative log of this:

$$
\begin{aligned}
&-\log(P(X_1 \in [x_1, x_1+\epsilon], X_2 \in [x_2, x_2+\epsilon], \ldots, X_N \in [x_N, x_N+\epsilon] \mid \boldsymbol{\theta})) \\
\approx &-N\log(\epsilon) - \sum_i \log(p(x_i \mid \boldsymbol{\theta})).
\end{aligned}
$$

- Again, $-Nlog(\epsilon)$ does not depend on $\boldsymbol{\theta}$
- We only need to optimize $-\sum_i \log(p(x_i \mid \boldsymbol{\theta})).$