Machine Learning - 20/9/2022

Probability and Statistics refresher

Basic properties of probability

Basic definitions

$$P(X) \geq 0, P(X) \in [0,1], \sum_{X \in \mathcal{X}} P(X) = 1$$

where $\mathcal{X}$ is our event universe and in the third property disjoint events were assumed.

Disjoint means $A \cap B = \emptyset$. In general, we have:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This was done in the discrete case, but is easily generalized to continous random variables:

$$p(x) \geq 0, p(x) \in [0,1], \int_{\mathcal{X}} dx\, p(x) = 1$$

We can also describe multiple random variables through the concept of joint probability $P(A, B)$ i.e the probability of A, B occuring at the same time.

Of course, we can derive the marginal probability by summing over the second variable:

$$P(X) = \sum_{Y \in \mathcal{Y}} P(X, Y)$$

And again, we can bring these concepts to the continuous case:

$$p(x) = \int_{\mathcal{Y}} dy\, p(x, y)$$

This are really intuitive definitions since we are basically asking "what is the probability of observing x, no matter the value of y ?"

We also have the concept of conditional probability $P(X|Y)$, which should read "the probability of observing X given that we observed Y"

This leads to the rule:

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

... which readily generalizes to the continous case.

If $P(X,Y) = P(X)P(Y)$ holds, then X, Y are said to be statistically independent.

This is easy to understand since it is basically saying that $P(X|Y) = P(X)$, i.e "the probability of observing X does not depend on Y happenning"

Bayes' Theorem

This theorem is the basis of the Bayesian formulation of probability.

We assume a variable θ follows a subjectively chosen distribution $\pi(\theta)$, called the **prior distribution**.

Then, we consider observations of samples of x, a random variable that is somehow dependent on θ but actually follows a distribution $f(x,\theta) = f(x|\theta)\pi(\theta)$.

These observations allows us to **update our belief** about the distribution of θ by applying Bayes' Theorem:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int d\theta\, f(x,\theta)\pi(\theta)}$$

This updated distribution $\pi(\theta|x)$ is called the **posterior distribution**.

Bayes' theorem holds for discrete variables:

$$P(X|Y) = \frac{P(Y|X)P(X)}{\sum_i P(X, Y_i)}$$

or simplifying even further, if $X \cup Y = \mathcal{U}$:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

More on continous random variables

If X is a continuous random variable, we can describe its statistical distribution through the **cumulative distribution function (CDF)**:

$$F_X(x) = P(X \leq x)$$

and one can describe the probability of observing $a < x < b$ as $F_X(b) - F_X(a)$.

This leads to the **probability density distribution (PDF)**:

$$p_X(x) = \frac{dF_X(x)}{dx} \Leftrightarrow F_X(x) = \int_{-\infty}^{x} dz\, p_X(z)$$

and naturally the probability of observing $a < x < b$ is simply $\int_a^b dx\, p_X(x)$.

Of course $\int_{-\infty}^{+\infty} dx\, p_X(x) = 1$.

Define the **mean** and **variance** as:

$$\mathbf{E}(x) = \int_{-\infty}^{+\infty} dx\, xp(x), \sigma_x^2 = \int_{-\infty}^{+\infty} dx\, (x - \mathbf{E}(x))^2 p(x)$$

If we have a function $f(x)$:

$$\mathbf{E}(f(x)) = \int_{-\infty}^{+\infty} dx\, f(x)p(x)$$

Discrete case is obtained simply changing integrals for summations.

Generalize to multivariate case:

$$\mathbf{E}(x, y) = \mathbf{E}_x(\mathbf{E}_{x|y}(f(x, y)))$$

Also introduce **covariance**:

$$\mathrm{cov}(x, y) = \mathbf{E}[\,(x - \mathbf{E}[x])(y - \mathbf{E}[y])\,]$$

and **correlation**:

$$r_{x,y} = \mathbf{E}[x, y] = \mathrm{cov}(x, y) - \mathbf{E}[x]\mathbf{E}[y]$$

Generalize even further introducing random vectors $\mathbf{x} \in \mathbf{R}^l$, **think about it as a column vector**.

The **covariance matrix** is given by:

$$\mathrm{Cov}(\mathbf{x}) = \mathbf{E}[\,(\mathbf{x} - \mathbf{E}[\mathbf{x}])(\mathbf{x} - \mathbf{E}[\mathbf{x}])^T\,]$$

i.e each entry $C_{i,j} = \mathrm{cov}(x_i, x_j)$

The **correlation matrix** is given by:

$$R_x = \mathbf{E}[\mathbf{x}\mathbf{x}^T]$$

i.e each entry $R_{i,j} = \mathbf{E}[x_i x_j]$

Relationship between them:

$$R_x = \text{Cov}(\mathbf{x}) + \mathbf{E}[\mathbf{x}]\mathbf{E}[\mathbf{x}^T]$$

They are positive semidefinite, i.e

$$\mathbf{y}^T \mathbf{A} \mathbf{y} \geq 0, \qquad \forall \mathbf{y} \in \mathbf{R}^l$$

An useful blog post about random vectors/matrices.

Note: application of expectation to random vector is done element-wise

Important distributions

1. Bernoulli

For binary random variables $x \in [0, 1]$ with $P(x = 1) = p$ and $P(x = 0) = 1 - p$.

$$P(x) = p^x (1 - p)^{1-x}$$

It simple to see that $\mathbf{E}[x] = p$ and $\sigma_x^2 = p(1 - p)$

2. Binomial

Defined for $x \in [0, 1, \ldots, n]$, given by:

$$P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

It simple to see that $\mathbf{E}[x] = np$ and $\sigma_x^2 = np(1 - p)$

3. Gaussian

Defined for $x \in \mathbf{R}$, parametrized by μ, σ^2 which coincide with its expectation and variance.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

4. Multivariate Gaussian

For random vector $\mathbf{x} \in \mathbf{R}^l$, written $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma})$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{l/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)\right)$$

This time, we have $\mathbf{E}[\mathbf{x}] = \mu$ and $\text{Cov}(\mathbf{x}) = \mathbf{\Sigma}$.

Important results:

The curves defined by $p(\mathbf{x}) = const.$ are hyper-ellipsoids, whose axis coincide with the directions of $\mathbf{\Sigma}$'s eigenvalues.

If x_i are statistically independent, then we have $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_l^2)$, **but this only holds for this distribution! Uncorrelated does not imply independence in general!**

5. Multinomial

Generalize binomial for the case where variable is not binary, but can take K possible values x_k with probability P_k.

There are two constraints:

$$\sum_k x_k = n, \sum_k P_k = 1$$

The distribution is given by:

$$P(x) = \binom{n}{x_1, x_2, \dots, x_K} \prod_k P_k^{x_k}$$

6. Mixtures of random variables

Consider X_1, X_2 with different distributions.

- Case #1

$$Z = X_1 + C$$

Z is distributed the same way as X_1 but takes a different range

- Case #2

$$Z = X_1 + X_2$$

This is a sum of random variables.

Z distribution changes. It can be shown that it is given by the convolution of the distributions of X_1, X_2

- Case #3

$$P(Z = z) = 0.3P(X_1 = z) + 0.7P(X_2 = z)$$

This is a **mixture of random variables**

The weights are interpreted as the probability of Z being sampled from each distribuition.

A little generalization:

$$\begin{cases} P(Z = z) = \sum_i w_i P(X_i = z), \\ \sum_i w_i = 1 \end{cases}$$

e.g A mixture of Gaussians is usually useful to describe multimodal populations.