Machine learning - Probability; Univariate

Yonghoon Dong

March 8, 2024

What is probability?

There are actually two different interpretations of probability.

1. Frequentist : probabilities represent long run **frequencies of event** that can happen multiple times
2. Bayesian : probability is used to quantify our **uncertainty** about something; hence it is fundamentally related to the information rather than repeated trials

One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about one-off events that don't have long term frequencies.

Types of uncertainty

The uncertainty in our predictions can arise for two fundamentally different reasons.

1. model uncertainty : our ignorance of the underlying hidden causes or mechanism generating our data
2. data uncertainty : arises from intrinsic variability, which cannot be reduced even if we collect more data

Mode

the mode of a distribution is the value with the highest probability mass or probability density

$$x^* = \arg\max_x p(x)$$

If the distribution is **multimodal**, this may not be unique.

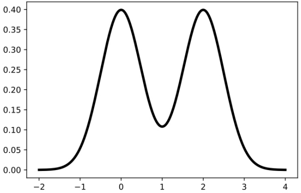


Figure: Illustration of a multimodal distribution

Law of total expectation (law of iterated expectations)

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$$

Proof.

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}[X|Y]] &= \mathbb{E}_Y\left[\sum_x x \cdot p(X = x|Y]\right] \\
&= \sum_y \left[\sum_x x \cdot p(X = x|Y = y)\right] p(Y = y) \\
&= \sum_{x,y} x \cdot p(X = x, Y = y) = \mathbb{E}[X]
\end{aligned}$$

□

Law of total variation (conditional variance)

$$\mathbb{V}[X] = \mathbb{E}_Y[\mathbb{V}[X|Y]] + \mathbb{V}[\mathbb{E}[X|Y]]$$

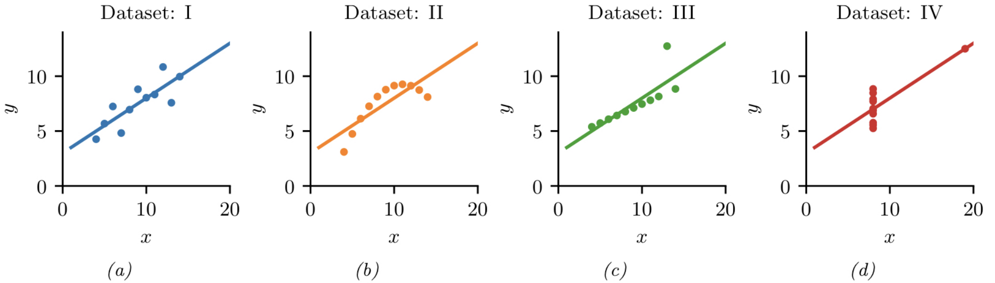
Proof.

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

□

Law of total variation (conditional variance)

Consider a mixture of K univariate Gaussians. Let Y be the hidden indicator variable that specifies which mixture component we are using, and let $X = \sum_{y=1}^{K} \pi_y \mathcal{N}(X|\mu_y, \sigma_y)$



(a) *(b)* *(c)* *(d)*

We get the intuitive result that the variance of X is dominated by which centroid it is drawn from, rather than the local variance around each centroid.

Bayes' rule

We will discuss the basics of **Bayesian inference**.

1. Inference : calculated degrees of certainty
2. Bayesian : inference methods that represent "degree of certainty" using probability theory.

Bayes' rule is just a formula for computing the probability distribution over possible values of an **unknown quantity** H given some **observed data** $Y = y$

$$p(H = h|Y = y) = \frac{p(H = h)p(Y = y|H = h)}{p(Y = y)}$$

Bayes' rule

1. p(H) : represents what we know about possible values of H before we wee any data, it is called the **prior distribution**
2. $p(Y|H = h)$: represents the distribution over the possible outcomes Y we expect to see if $H = h$, it is called the **observation distribution**
3. $p(Y = y|H = h)$: when we evaluate this at a point corresponding to the actual observations, y, we get the function $p(Y = y|H = h)$, it is called the **likelihood** (Note that this is a function of h)
4. $p(H = h|Y = y)$: normalizing the joint distribution by computing $p(H = h, Y = y)/p(Y = y)$ for each h gives the **posterior distribution** $p(H = h|Y = y)$; this represents our new **belief state** about the possible values of H

Bayes' rule

We can use the symbol $\propto$ to denote proportional to, since we are ignoring the denominator, which is just a constant, independent of H.

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Using Bayes rule to update a distribution over unknown values of some quantity of interest, given relevant observed data, is called **Bayesian inference**.

Bernoulli distribution

Consider tossing a coin, where the probability of event that is lands heads is given by $0 \leq \theta \leq 1$.

1. $Y = 1$: coin lands heads
2. $Y = 0$: coin lands tails

Thus we are assume that

$$
\begin{aligned}
p(Y = 1) &= \theta \\
p(Y = 0) &= 1 - \theta
\end{aligned}
$$

This is called the **Bernoulli distribution**, and can be written as

$$Y \sim \mathrm{Ber}(\theta)$$

where the symbol $\sim$ means "is sampled from" or "is distributed as"

Bernoulli distribution

The probability mass function of this distribution is defined as follows

$$\text{Ber}(y|\theta) = \begin{cases} 1-\theta & \text{if } y = 0 \\ \theta & \text{if } y = 1 \end{cases}$$

We can write this in more concise manner as follows

$$\text{Ber}(y|\theta) = \theta^{y}(1-\theta)^{1-y}$$

Binomial distribution

The Bernoulli distribution is a special case of the **binomial distribution**. We can view the binomial as a set of independent N Bernoulli trials.

$$\mathrm{Bin}(s|N,\theta) = \binom{n}{k} \theta^s (1-\theta)^{N-s}$$

Binary classification problem : sigmoid (logistic) function

When we want to predict a binary variable $y \in \{0, 1\}$ given some input x, we need to use a **conditional probability distribution** of the form

$$p(y|x, \theta) = \mathrm{Ber}(y|f(x; \theta))$$

where $f(x; \theta)$ is some function that predicts the parameter in the Bernoulli.

Binary classification problem : sigmoid (logistic) function

To avoid the requirement of $0 \leq f(x;\theta) \leq 1$, we can use the following model and assume f be an unconstrained function

$$p(y|x,\theta) = \mathrm{Ber}(y|\sigma(f(x;\theta))$$

where θ is the logistic (sigmoid) function, defined as

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Binary classification problem : sigmoid (logistic) function

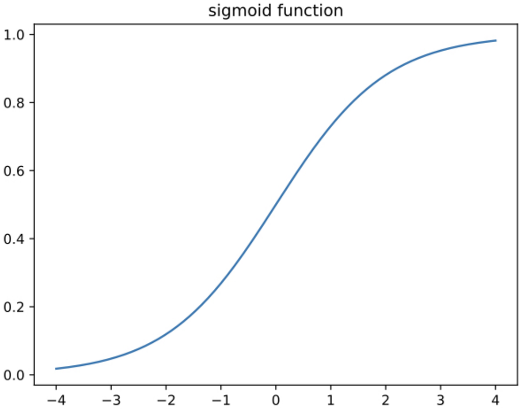


Figure: Illustration of a logistic (sigmoid) function

Example : Binary logistic regression

In binary logistic regression, we use

$$f(x;\theta) = w^T x + b$$

Thus the model has the form

$$p(y|x;\theta) = \mathrm{Ber}(y|\sigma(w^T x + b))$$

In other words,

$$p(y=1|x;\theta) = \sigma(w^T x + b) = \frac{1}{1+e^{-(w^T x + b)}}$$

Categorical classification problem : softmax function

To represent a distribution over a finite set of labels, $y \in \{1, \dots, C\}$, we can use the **categorical distribution**, which generalizes the Bernoulli distribution. The categorical distribution is a discrete probability distribution with one parameter per class

$$\text{Cat}(y|\theta) = \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)}$$

In otherwords,

$$p(y = c|\theta) = \theta_c$$

The categorical distribution is a special case of the **multinomial distribution**.

Categorical classification problem : softmax function

We can define

$$p(y|x,\theta) = \mathrm{Cat}(y|f(x;\theta))$$

Similar to the binary case, we want to avoid the requirement of f. It is common to pass the output from f into **softmax** function, also called the **multinomial logit**. This is defined as follows

$$\mathrm{softmax}(a) = \left[\frac{e^{a_1}}{\sum_{c'=1}^{C} e^{a_{c'}}} \quad \cdots \quad \frac{e^{a_C}}{\sum_{c'=1}^{C} e^{a_{c'}}}\right]$$

where the inputs to the softmax, $a = f(x;\theta)$ are called **logits**

Example : Categorical logistic regression

If we use a linear predictor of the form

$$f(x;\theta) = Wx + b$$

where W is a $C \times D$ matrix, and b is a C-dimensional bias vector. Therefore,

$$p(y|x;\theta) = \mathrm{Cat}(y|\mathrm{softmax}(Wx + b))$$

Let $a = Wx + b$ be the logits, then

$$p(y = c|x;\theta) = \frac{e^{a_c}}{\sum_{c'=1}^{C} e^{a_{c'}}}$$

This is known as **multinomial logistic regression**.

Connection of logistic and categorical logistic regression

If we have just two classes, this reduces to binary logistic regression.

$$\text{softmax}(a)_0 = \frac{e^{a_0}}{e^{a_0} + e^{a_1}} = \frac{1}{1 + e^{a_1 - a_0}} = \sigma(a_0 - a_1)$$

Therefore we can interpret this model is **over-parameterized**.

Regression with Gaussian distribution

In some cases, it is helpful to make the parameters of the Gaussian be functions of some input variables, i.e. we want to create a conditional density model as follows

$$p(y|x;\theta) = \mathcal{N}(y|f_\mu(x;\theta), f_\sigma(x;\theta)^2)$$

where $f_\mu(x;\theta)$ predicts the means, and $f_\sigma(x;\theta)^2$ predicts the variance.

Regression with Gaussian distribution

1. It is common to assume that the variance is fixed, and is independent of the input. This is called **homoscedastic regression**.
2. Furthermore, it is common to assume the mean is a linear function of the input. The resulting model is called **linear regression**. Therefore, a conditional density model is

$$p(y|x;\theta) = \mathcal{N}(y|w^T x + b, \sigma^2)$$

where $\theta = (w, b, \sigma^2)$

Regression with Gaussian distribution

We can also make the variance depend on the input; this is called **heteroschedastic regression**. Then we have

$$p(y|x;\theta) = \mathcal{N}(y|w_\mu^T x + b, \sigma_+(w_\sigma^T x))$$

where $\theta = (w_\mu, b, w_\sigma)$, and

$$\sigma_+(a) = \log(1 + e^a)$$

is called **softplus** function.

Why Gaussian distribution so widely used?

The Gaussian distribution is the most widely used distribution in statistics and machine learning.

1. It has two parameters which are easy to interpret, and which capture some of the most basic properties of a distribution
2. The central limit theorem tells us that sums of independent random variables have an approximately Gaussian distribution, making it a good choice for modeling residual errors or noise.
3. It has the least number of assumptions subject to the constraint of having a specified mean and variance
4. It has a simple mathematical form

Empirical distribution

Suppose we have a set of N samples $D = \{x^{(1)}, \ldots, x^{(N)}\}$, derived from a distribution $p(X)$. We can approximate the pdf using a set of delta functions,

$$\hat{p}_N(x) = \frac{1}{N} \sum_{n=1}^{N} \delta_{x^{(n)}}(x)$$

This is called the **empirical distribution** of the dataset $\mathcal{D}$. The corresponding cdf is given by

$$\hat{P}_N(x) = \frac{1}{N} \sum_{n=1}^{N} u_{x^{(n)}}(x)$$

where $u_y(x)$ is a **step function** at y

Biliography

1. Probabilistic Machine Learning: An introduction, Kevin P. Murphy