Introduction: Basic Probability and Statistics

Introduction to Artificial Intelligence with Mathematics
Lecture Notes

Ganguk Hwang

Department of Mathematical Sciences
KAIST

Basic Probability

The probability $P\{E\}$ of an event E

$N(E;n) =$ the number of occurrences of the event E during n experiments

$n =$ the total number of experiments

$$P\{E\} = \lim_{n\to\infty} \frac{N(E;n)}{n}.$$

Axioms of probability

- $0 \le P\{E\} \le 1, \qquad E \subset \Omega$
- For the sample space Ω, $P\{\Omega\} = 1$
- If E_i are mutually disjoint, i.e., $E_i \cap E_j = \phi$ for $i \neq j$,

$$P\{\cup_{i=1}^{\infty} E_i\} = \sum_{i=1}^{\infty} P\{E_i\}$$

Properties of probability

- $P\{\phi\} = 0$
- $P\{E^c\} = 1 - P\{E\}$ where $E^c = \Omega - E$
- $P\{E \cup F\} = P\{E\} + P\{F\} - P\{E \cap F\}$
- If $E \subseteq F$, then $P\{E\} \leq P\{F\}$
- For events $\{E_1, E_2, \cdots\}$ with $E_i \subset E_{i+1}, i = 1, 2, \cdots$

$$P\{\cup_{i=1}^{\infty} E_i\} = \lim_{n \to \infty} P\{E_n\}.$$

- For events $\{E_1, E_2, \cdots\}$ with $E_i \supset E_{i+1}, i = 1, 2, \cdots$

$$P\{\cap_{i=1}^{\infty} E_i\} = \lim_{n \to \infty} P\{E_n\}.$$

Conditional Probability

The conditional probability $P\{F|E\}$ of event F, given event E, is defined by

$$P\{F|E\} = \frac{P\{E \cap F\}}{P\{E\}}, P\{E\} > 0.$$

Law of total probability

Let $\{E_i\}_{i=1}^N$ be a partition of the sample space Ω, i.e., $\cup_{i=1}^N E_i = \Omega$ and E_i are mutually disjoint. Then

$$P\{F\} = \sum_{i=1}^{N} P\{E_i\}P\{F|E_i\}.$$

Note that N may be either finite or infinite, and typically $P\{E_i\}$ is easy to calculate.

Bayes' formula

For a partition $\{E_i\}$ of the sample space Ω and an event F,

$$P\{E_i|F\} = \frac{P\{E_i \cap F\}}{P\{F\}} = \frac{P\{F|E_i\}P\{E_i\}}{\sum_{i=1}^{N} P\{E_i\}P\{F|E_i\}}$$

Independence

Two events E and F are independent if $P\{E \cap F\} = P\{E\}P\{F\}$.
Consequently, we have $P\{F|E\} = P\{F\}$.
Example. Tossing two coins

- Sample space $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$
- $E =$ the first coin lands on head, $F =$ the second coin lands on head
- $P\{E\} = \frac{1}{2}, P\{F\} = \frac{1}{2}$
- $P\{F|E\} = \frac{P\{E \cap F\}}{P\{E\}} = \frac{1/4}{1/2} = \frac{1}{2}$

Random variable

A random variable X is a real valued mapping from the sample space Ω to the set $\mathbb{R}$ of real numbers, which associates a real number to each element $\omega \in \Omega$.

Example. When tossing a coin n times, the number of heads in n tosses is denote by a random variable X with state space $\{0, 1, 2, \cdots, n\}$.

Distribution

For a random variable (R.V.) X, its distribution function $F(x)$ (or $F_X(x)$) is defined by $F(x) = P\{X \leq x\}$ ($X \sim F(x)$). Sometimes it is also called cumulative distribution function (cdf).

Complementary distribution function (tail distribution)

$$G(x) = 1 - F(x) = P\{X > x\}$$

Function of random variable Let X be a random variable and $g(x)$ be a strictly increasing function from $\mathbb{R}$ to $\mathbb{R}$.
Consider a random variable $Y = g(X)$. Observe that

$$\{Y \leq y\} = \{g(X) \leq y\} = \{X \leq g^{-1}(y)\}.$$

Then we have

$$F_Y(y) = F_X(g^{-1}(y)).$$

For the special case $g(x) = F_X(x)$ (consequently, $0 \leq g(x) \leq 1$), we have

$$F_Y(y) = F_X(F_X^{-1}(y)) = y, \qquad 0 \leq y \leq 1,$$

i.e., $Y = F_X(X) = U$ or $X = F_X^{-1}(U)$ where U is the uniform R.V. in $[0, 1]$.
In simulation, if one can generate random numbers in $[0, 1]$, then we can draw values for an arbitrary random variable X.

Expectation

The expectation $E[X]$ of a R.V. X is defined by

- Continuous R.V. : $E[X] = \int_{-\infty}^{\infty} xf(x)dx$, where $f(x)$ is the density function of X, i.e., $f(x) = \frac{d}{dx}F_X(x)$.
- Discrete R.V. : $E[X] = \sum_n np_n$ where $p_n = P\{X = n\}$

When X is a nonnegative R.V., we have

- Continuous R.V. :

$$\begin{aligned} E[X] &= \int_0^{\infty} P\{X > x\}dx = \int_0^{\infty}\int_x^{\infty} f(y)dydx \\ &= \int_0^{\infty}\int_0^{y} f(y)dxdy = \int_0^{\infty} yf(y)dy. \end{aligned}$$

- Discrete R.V. : $E[X] = \sum_{n=1}^{\infty} P\{X \geq n\}$.

Let X_i be independent and identically distributed (i.i.d.) R.V.s with distribution $F(x)$.

Distribution of the maximum

$$\begin{aligned} P\{\max(X_1, X_2, \cdots, X_n) \leq x\} &= P\{X_1 \leq x, \cdots, X_n \leq x\} \\ &= P\{X_1 \leq x\} \cdots P\{X_n \leq x\} \\ &= F(x)^n. \end{aligned}$$

Distribution of the minimum

$$\begin{aligned} P\{\min(X_1, X_2, \cdots, X_n) > x\} &= P\{X_1 > x, \cdots, X_n > x\} \\ &= P\{X_1 > x\} \cdots P\{X_n > x\} \\ &= (1 - F(x))^n. \end{aligned}$$

Change of Variables

Let U be an open set in $\mathbb{R}^n$ and

$$g = (g_1, \cdots, g_n) : U \to \mathbb{R}^n$$

be a one-to-one function on U. For random vectors $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_n)$, let $Y = g(X)$, i.e., $Y_i = g_i(X_1, \cdots, X_n)$. Assume that the coordinate functions g_i are continuously differentiable on U. In addition, for the Jacobian matrix of g, defined by

$$J_g(x_1, \cdots, x_n) = (\frac{\partial g_i}{\partial x_j})_{1 \leq i,j \leq n},$$

assume that $|\det[J_g(x_1, \cdots, x_n)]| > 0$ on U.

Then, by the inverse function theorem for $V = g(U)$ there exists an inverse function

$$g^{-1} : V \to \mathbb{R}^n$$

with $|\det[J_{g^{-1}}(y_1, \cdots, y_n)]| > 0$ on V and

$$\det[J_{g^{-1}}(y_1, \cdots, y_n)] = \frac{1}{\det[J_g(g^{-1}(y_1, \cdots, y_n))]}.$$

Furthermore, by using the change of variables, when the joint probability density function of X is given by $f_X(x_1, \cdots, x_n)$, the joint probability density function of Y is given by

$$f_Y(y_1, \cdots, y_n) = f_X(g^{-1}(y_1, \cdots, y_n)) \frac{1}{|det[J_g(g^{-1}(y_1, \cdots, y_n))]|}.$$

The proof is as follows: For a subset $A \in V$ and $Y = (y_1, \cdots, y_n)$,

$$\begin{aligned} & P\{Y \in A\} \\ & = P\{X \in g^{-1}(A)\} \\ & = \int \cdots \int_{(x_1, \cdots, x_n) \in g^{-1}(A)} f_X(x_1, \cdots, x_n) dx_1 \ \cdots \ dx_n \\ & = \int \cdots \int_A f_X(g^{-1}(y_1, \cdots, y_n)) \frac{1}{|det[J_g(g^{-1}(y_1, \cdots, y_n))]|} dy_1 \ \cdots \ dy_n. \end{aligned}$$

Example: Let X_1 and X_2 be independent uniform R.V.s on $[0, 1]$, and

$$Y_1 = X_1 + X_2, Y_2 = X_1 - X_2.$$

Find the joint probability density function of Y_1 and Y_2.

Let $y_1 = g_1(x_1, x_2) = x_1 + x_2$ and $y_2 = g_2(x_1 + x_2) = x_1 - x_2$. Then the determinant of the jacobian matrix is

$$\begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2.$$

For $0 \leq x_1, x_2 \leq 1$, $f_{X_1,X_2}(x_1, x_2) = 1$. In addition, we get $0 \leq y_1 + y_2 \leq 2$ and $0 \leq y_1 - y_2 \leq 2$. Hence, it follows that

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2} I_{\{0 \leq y_1 + y_2 \leq 2, 0 \leq y_1 - y_2 \leq 2\}}(y_1, y_2).$$

Convex Function

A real-valued function f is called convex if, for any x, y and $t \in [0, 1]$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Jensen's Inequality

Let X be a random variable and f be a convex function. Then

$$f(E[X]) \leq E[f(X)].$$

Basic Bayesian Statistics

Bayesian Inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.

Bayesian updating is particularly important in the dynamic analysis of a sequence of data.

Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, sport, and law.

(in Wikipedia)

Thomas Bayes and Bayes' Theorem

Thomas Bayes (1701 ĩ1761) was an English statistician, philosopher and Presbyterian minister who is known for formulating a specific case of the theorem that bears his name: Bayes' theorem.

Bayes never published what would become his most famous accomplishment; his notes were edited and published after his death by Richard Price.

(in Wikipedia)

Bayes' Theorem

For a partition $\{E_i\}$ of the sample space Ω and an event F,

$$P\{E_i|F\} = \frac{P\{E_i \cap F\}}{P\{F\}} = \frac{P\{F|E_i\}P\{E_i\}}{\sum_{i=1}^{N} P\{E_i\}P\{F|E_i\}}$$

Distribution Function and Parameter

For the *distribution function* $F(x)$ of a random variable X, i.e.,

$$F(x) = P\{X \leq x\},$$

the *parameter* of a distribution function is the value that determines the distribution.

For instance, when we consider a Bernoulli random variable X with $P\{X = 1\} = p, P\{X = 0\} = 1 - p$, the value p determines the distribution of X and is the parameter of the distribution of X.

Likelihood Function

Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed (i.i.d.) random variables with parameter θ. The probability mass (or density) function of X_i is given by $p(x|\theta)$. Then, the likelihood $\mathcal{L}(x_1, x_2, \cdots, x_n|\theta)$ is defined by

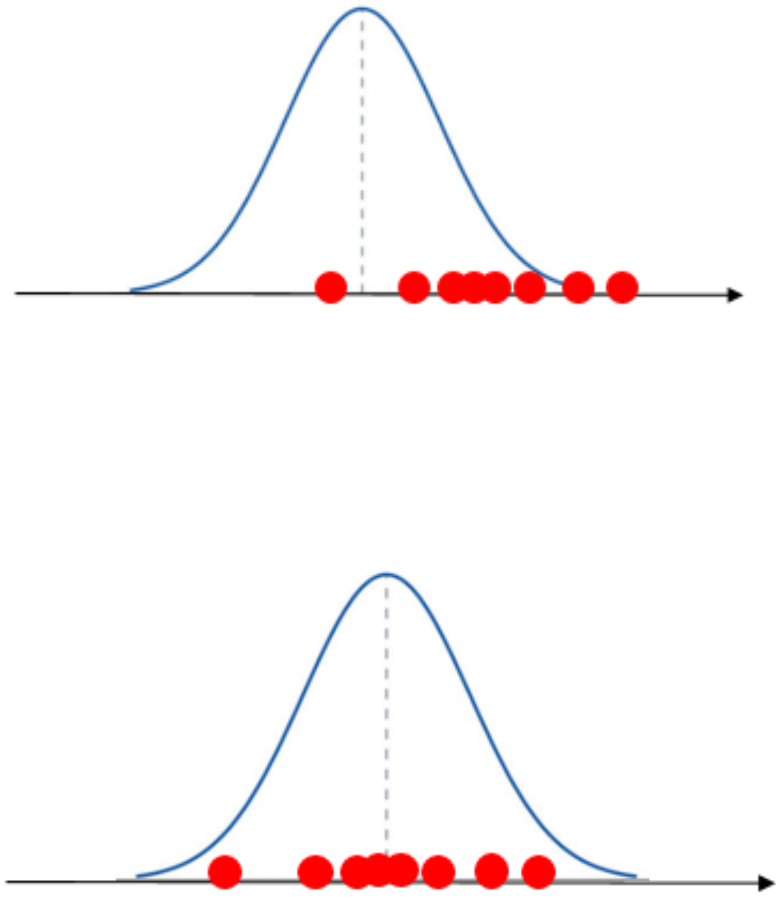
$$\mathcal{L}(x_1, x_2, \cdots, x_n|\theta) = p(x_1|\theta)p(x_2|\theta)\cdots p(x_n|\theta)$$

Maximum Likelihood Estimator (MLE) of θ

- The MLE is the value of θ that maximizes the likelihood function.
- Here, we consider the likelihood function as a function of θ.
- To compute the MLE, we usually use $\log(\mathcal{L}(x_1, x_2, \cdots, x_n|\theta))$.

The concept of MLE

Given a data set, which one does explain the data set better?



MLE: an example

Let $X_1, X_2, \cdots, X_n$ be indepedent and identically distributed (i.i.d.) Bernoulli random variables where

$$P\{X_1 = 1|\theta\} = \theta, \qquad P\{X_1 = 0|\theta\} = 1 - \theta.$$

Here, θ is usually called the parameter of Bernoulli distribution.

First, consider the likelihood function of a given data set $\{x_1, x_2, \cdots, x_n\}$.

$$\begin{aligned} & P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta\} \\ &= \prod_{i=1}^{n} P\{X_i = x_i|\theta\} = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}. \end{aligned}$$

To find the MLE, we use

$$\begin{aligned}&\log(P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta\}) \\ &= \left(\sum_{i=1}^{n} x_i\right)\log(\theta) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta)\end{aligned}$$

which is considered as a function of θ.

By differentiating $\log(P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta\})$ with respect to θ and letting it be 0, the MLE $\hat{\theta}$ is given by

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Bayesian models

In Bayesian models we use the Bayes' rule to obtain unknown probability.

In Bayesian models with two random varaibles X and Y, the following are initially given.

- The data generating distribution: $X|Y \sim p(x|y)$
- The prior distribution $Y \sim p(y)$

We then obtain a sample value $X = x$ and want to estimate the distribution (the posterior distribution) of Y, given $X = x$, i.e., $p(y|x)$.

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x,y)\ dy}.$$

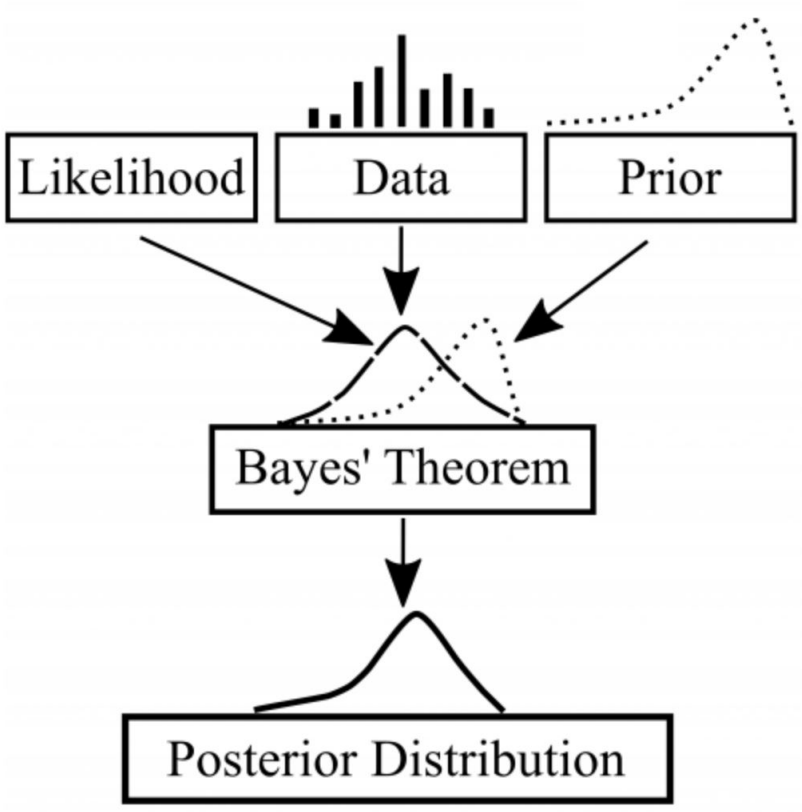
We frequently use $p(y|x) \propto p(x|y)p(y)$.

Consider a random variable X having distribution function $F(x)$ with unknown parameter θ.

In Bayesian models, the unknown parameter θ is considered *stochastic*. So we believe that $\theta \sim p(\theta)$ where $p(\theta)$ is called a *prior* distribution before sampling. After sampling, using Bayes' rule we obtain so-called a *posterior* distribution as follows:

$$p(\theta|x) = \frac{p(x,\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x,\theta)\ d\theta}.$$

- $P(\theta)$ is the *prior* which is our belief of θ without considering the data (evidence) $\mathcal{D}$
- $P(\theta|\mathcal{D})$ is the *posterior* which is a refined belief of θ with the evidence $\mathcal{D}$
- $P(\mathcal{D}|\theta)$ is the *likelihood* which is the probability of obtaining the data $\mathcal{D}$ when generated with parameter θ
- $P(\mathcal{D})$ is the *evidence* which is the probability of obtaining the data by considering all possible values of θ



Courtesy: http://jason-doll.com/wordpress/?page_id=127

Figure: Concept of Bayesian models

Maximum A Posterior Estimator

Consider $X \sim p(x|\theta)$ with a prior distribution of $p(\theta)$. Here, θ is the parameter of the distribution of X.

The Maximum A Posterior (MAP) estimator is given by

$$\theta_{MAP} = \text{argmax}_\theta \log p(\theta|X).$$

Note that $p(\theta|X)$ is the posterior distribution of θ and

$$p(\theta|X) = \frac{p(\theta, X)}{p(X)}.$$

Since $p(X)$ is not a function of θ, we have

$$\theta_{MAP} = \text{argmax}_\theta \log p(\theta, X).$$

Recall that the Maximum Likelihood Estimator (MLE) is defined by

$$\theta_{MLE} = \text{argmax}_\theta \log p(X|\theta)$$

and the MAP estimator is given by

$$\theta_{MAP} = \text{argmax}_\theta \log p(\theta, X) = \text{argmax}_\theta \log p(X|\theta)p(\theta).$$

So, the MAP estimator can use the prior information, but the MLE cannot.

c.f. Frequentist vs. Bayesian
The MAP is the Bayesian estimator while the MLE is the frequentist estimator.

Bayesain model: an example

Let $X_1, X_2, \cdots, X_n$ be indepedent and identically distributed (i.i.d.) Bernoulli random variables where

$$P\{X_1 = 1|\theta\} = \theta, \qquad P\{X_1 = 0|\theta\} = 1 - \theta.$$

Here, θ is usually called the parameter of Bernoulli distribution. First, observe that

$$\begin{aligned} & P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta\} \\ &= \prod_{i=1}^{n} P\{X_i = x_i|\theta\} = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}. \end{aligned}$$

The prior distribution of θ is given by a uniform distribution over $[0, 1]$, i.e.,

$$p(\theta) = 1 \text{ for } 0 \leq \theta \leq 1.$$

After obtaining sample values $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, we are interested in the posterior distribution of θ.

$$\begin{aligned} & p(\theta | X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ & \quad \propto P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n | \theta) p(\theta) \\ & \quad \propto \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}. \end{aligned}$$

Considering the normalization constant, we obtain

$$p(\theta|X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$
$$= \frac{\Gamma(n+2)}{\Gamma(1+\sum_{i=1}^n x_i)\Gamma(1+n-\sum_{i=1}^n x_i)}\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

which is a Beta distribution with parameters $a = \sum_{i=1}^n x_i + 1$ and $b = n - \sum_{i=1}^n x_i + 1$.

c.f.

$$\text{Beta}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

Note that, when $a = b = 1$, $\text{Beta}(a,b)$ is in fact a uniform distribution over $[0,1]$. That is, the prior and the posterior of θ are both Beta distributions.

Example 1.
Suppose we want to buy something from Amazon.com, and there are two sellers offering it for the same price. Seller 1 has 90 positive reviews and 10 negative reviews. Seller 2 has 2 positive reviews and 0 negative reviews. Who should we buy from? (from Machine Learning by K.P. Murphy)

Let θ_1 and θ_2 be the unknown reliabilities of the two sellers. Since we don't know much about them, we will endow them both with uniform priors, $\theta_i \sim \text{Beta}(1,1), i = 1, 2$. The posteriors are

$$p(\theta_1|\mathcal{D}_1) = \text{Beta}(91, 11), \qquad p(\theta_2|\mathcal{D}_2) = \text{Beta}(3, 1).$$

Hence,

$$\begin{aligned} P(\theta_1 > \theta_2|\mathcal{D}_1, \mathcal{D}_2) &= \int_0^1 \int_0^1 I_{\{\theta_1 > \theta_2\}} \text{Beta}(\theta_1|91, 11)\text{Beta}(\theta_2|3, 1)\ d\theta_1 d\theta_2 \\ &= 0.710. \end{aligned}$$

This concludes that we are better off buying from seller 1.

Example 2.

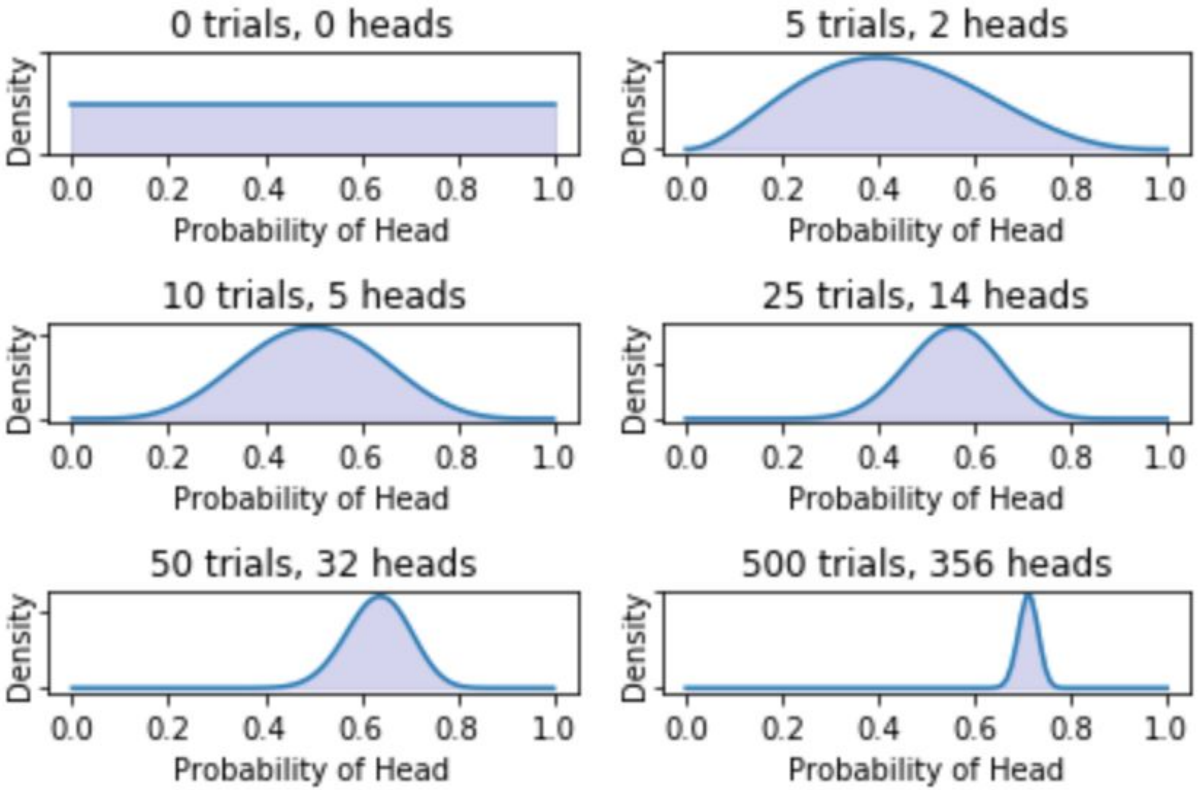


Figure: Bayesian Experiment for Bernoulli distribution with $p = 0.7$

More generally, if we use Beta(a, b) as a prior distribution of θ, we have

$$\begin{aligned} & p(\theta|X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ & \quad \propto P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta)p(\theta) \\ & \quad \propto P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ & \quad \propto \theta^{a+\sum_{i=1}^n x_i - 1}(1-\theta)^{b+n-\sum_{i=1}^n x_i - 1}. \end{aligned}$$

Hence, we see that

$$p(\theta|X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \text{Beta}(a + \sum_{i=1}^n x_i, b + n - \sum_{i=1}^n x_i).$$

In this case, the Beta distribution is a *conjugate prior*.

From now on, we use the notation $p(\theta|x_1, x_2, \cdots, x_n)$ or $p(\theta|X)$ for simplicity.

More on Conjugate Distributions

In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood function.

As shown before, if we use a conjugate prior, we can obtain a closed-form expression for the posterior. This also shows how the likelihood function updates a prior distribution.

Basic Information Theory

Entropy

- Consider a discrete random variable X with p.m.f. $p(x)$
- We want to define a measure $h(x)$ of the information of observing $X = x$.
- $h(x) = -\log p(x)$
  - If $p(x)$ is low (resp. high), $h(x)$ should be high (resp. low). So, $h(x)$ should be a monotonically decreasing function of $p(x)$.
  - If X and Y are independent, $h(x, y)$ should be $h(x) + h(y)$. Here, we use $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

Definition 1

The entropy $H(X)$ of X is defined by

$$H(X) = -E[\log p(X)] = E\left[\log\left(\frac{1}{p(X)}\right)\right].$$

Properties of entropy

- $H(X) \geq 0$ since $p(x) \in [0, 1]$.
- $H(X) = 0$ if there exists some x with $p(x) = 1$.

Conditional Entropy

Definition 2

For $(X, Y) \sim p(x, y)$, the conditional entropy is defined by

$$H(Y|X) = -E_{X,Y}[\log p(Y|X)].$$

For $H(X,Y) = -E_{X,Y}[\log p(X,Y)]$, the joint entropy of X and Y, we see that

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

since

$$\begin{aligned} H(X,Y) &= -E_{X,Y}[\log p(X,Y)] = -E_{X,Y}[\log p(Y|X)p(X)] \\ &= -E_{X,Y}[\log p(Y|X) + \log p(X)] \\ &= -E_{X,Y}[\log p(Y|X)] - E_{X,Y}[\log p(X)] \\ &= -E_{X,Y}[\log p(Y|X)] - E_X[\log p(X)] \\ &= H(Y|X) + H(X). \end{aligned}$$

Kullback-Leibler Divergence

Definition 3

The Kullback-Leibler (KL) divergence $KL(p\|q)$ between two probability mass (density) functions $p(x)$ and $q(x)$ is defined by

$$KL(p\|q) = E_{p(X)}\left[\log\left(\frac{p(X)}{q(X)}\right)\right].$$

Here, we have the following conventions: for $a, b > 0$

$$0\log\left(\frac{0}{0}\right) = 0, 0\log\left(\frac{0}{a}\right) = 0, b\log\left(\frac{b}{0}\right) = \infty.$$

Note that we can also define the KL divergence in a similar way for continuous distributions.

The KL divergence $KL(p\|q)$ is not a true distance because it is not symmetric and does not satisfy the triangular inequality.

For instance, consider two Bernoulli distributions with respective success probabilities r and s ($r \neq s$). Then

$$\begin{aligned} KL(p\|q) &= (1-r)\log\left(\frac{1-r}{1-s}\right) + r\log\left(\frac{r}{s}\right), \\ KL(q\|p) &= (1-s)\log\left(\frac{1-s}{1-r}\right) + s\log\left(\frac{s}{r}\right). \end{aligned}$$

Theorem 1

$$KL(p\|q) \geq 0$$

Moreover, $KL(p\|q) = 0$ *if and only if* $p(x) = q(x)$ *for all* x.

Proof: Let $E = \{x|p(x) > 0\}$. Then

$$\begin{aligned}
KL(p\|q) &= \sum_{x\in E} p(x)\log\left(\frac{p(x)}{q(x)}\right) = -\sum_{x\in E} p(x)\log\left(\frac{q(x)}{p(x)}\right) \\
&\geq -\log\left(\sum_{x\in E} p(x)\frac{q(x)}{p(x)}\right) \text{ by Jensen's inequality} \\
&= -\log\left(\sum_{x\in E} q(x)\right) \\
&\geq 0.
\end{aligned}$$

From the above derivation, we see that the equality holds if and only if $\frac{q(x)}{p(x)} = c$ for all possible values of x. In this case,

$$1 = \sum_x q(x) = c \sum_x p(x) = c$$

and hence $q(x) = p(x)$ for all x. $\square$

Mutual Information

Definition 4

For $(X, Y) \sim p(x, y)$, the mutual information of X and Y is defined by

$$I(X;Y) = KL(p(x,y)||p(x)p(y)) = E_{p(x,y)}\left[\log\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right]$$

Note that

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

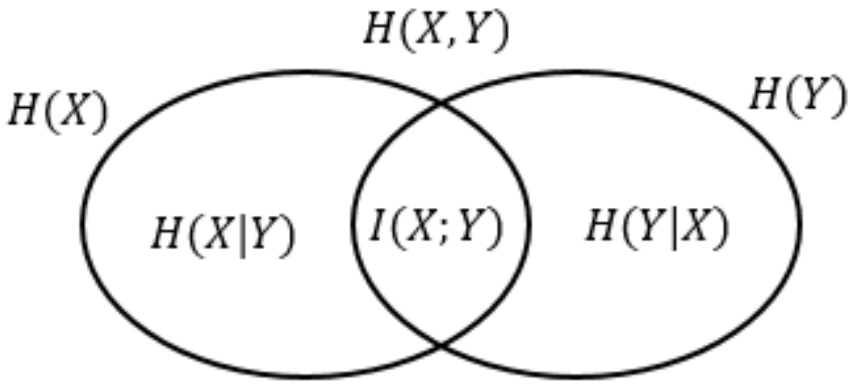
So $I(X;Y)$ is the reduction of the uncertainty on X obtained by telling the value of Y, and the reduction of the uncertainty on Y obtained by telling the value of X.

$$
\begin{aligned}
I(X;Y) &= E_{p(X,Y)}\left[\log\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right] \\
&= -E_{p(X,Y)}\left[\log\left(\frac{p(X)p(Y)}{p(X,Y)}\right)\right] \\
&= -E_{p(X,Y)}\left[\log\left(\frac{p(X)p(Y)}{p(Y|X)p(X)}\right)\right] \\
&= -E_{p(X,Y)}\left[\log\left(\frac{p(Y)}{p(Y|X)}\right)\right] \\
&= -E_{p(X,Y)}\left[\log\left(p(Y)\right)\right] + E_{p(X,Y)}\left[\log\left(p(Y|X)\right)\right] \\
&= H(Y) - H(Y|X).
\end{aligned}
$$

Similarly, we have

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

Entropy, Joint Entropy, Conditional Entropy, and Mutual Information



Cross Entropy

The cross entropy $H_p(q)$ of the distribution $q(x)$ relative to a distribution $p(x)$ is defined by

$$H_p(q) = -E_p[\log q(X)].$$

Note that

$$\begin{aligned} KL(p\|q) &= \sum_{x\in E} p(x) \log\left(\frac{p(x)}{q(x)}\right) \\ &= -\sum_{x\in E} p(x) \log\left(\frac{q(x)}{p(x)}\right) \\ &= -\sum_{x\in E} p(x) \log\left(q(x)\right) - \sum_{x\in E} p(x) \log\left(\frac{1}{p(x)}\right) \\ &= H_p(q) - H(p). \end{aligned}$$

So we see that $H_p(q) \geq H(p)$.