Machine Learning - Michaelmas Term 2016
Lecture 3 : Maximum Likelihood

Lecturer: Varun Kanade

In this lecture, we'll look at the linear model from a probabilistic perspective. Previously, we added a noise term ϵ to the linear model as a necessary correction for the case where the observed data does not exactly satisfy a linear relationship (which will almost always be the case). In this lecture, we see how the language of probability can be used to explicitly model this error term as noise generated from some distribution. We'll then develop tools to estimate the appropriate parameters for the model from this probabilistic perspective.

1 Probability Review

We'll very briefly discuss some concepts from probability theory below. For further details please refer to (Murphy, 2012, Chap 2).

1.1 Covariance and Correlation

Let X and Y be two real-valued random variables. The covariance is a measure of linear relationship between the two random variables. Formally, the covariance $\mathrm{cov}(X, Y)$ is defined as

$$\mathrm{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] \tag{1}$$

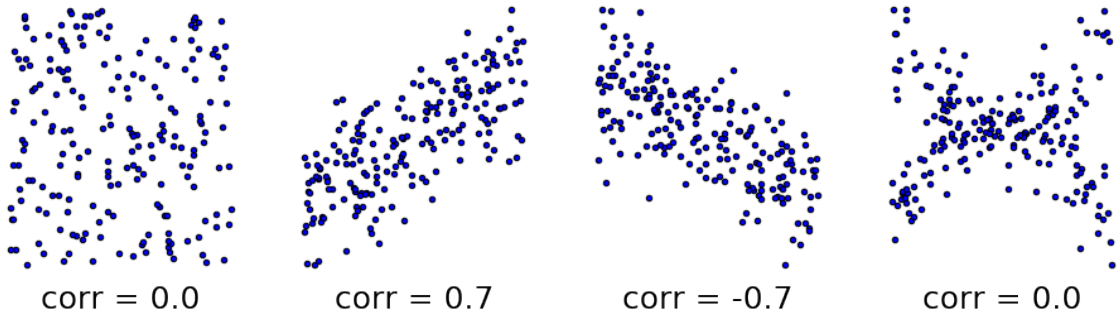
Covariance depends on the scale of the random variables. For example, if X measures the distance of a commute and Y the time required to undertake it, then if the distance is measured in metres as opposed to kilometres, then the covariance $\mathrm{cov}(X, Y)$ may vary by a factor of one thousand. The Pearson correlation coefficient normalises this so that the correlation between any two random variables is always between -1 and $+1$. The correlation coefficient $\mathrm{corr}(X, Y)$ is defined as follows:

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X) \cdot \mathrm{var}(Y)}} \tag{2}$$

Figure 1 shows different pairs of random variables (X, Y) and their correlation coefficients. Independent random variables indeed have 0 correlation as can be seen in Figure 1(a). However, the converse it not true. Random variables having 0 correlation can be very much dependent (see Fig. 1(d)). As an extreme example of this, if X is distributed uniformly on $[-1, 1]$ and $Y = X^2$, then the $\mathrm{cov}(X, Y) = 0$, however clearly X and Y are not independent!

For a random variable $\mathbf{x} \in \mathbb{R}^D$, the covariance matrix contains the covariance between all pairs of components. The diagonal terms contains the variance of each component. One can similarly write out a correlation matrix, which contains pairwise correlations instead of covariances; in this case the diagonal will have all entries equal to 1.

$$\mathrm{cov}(\mathbf{x}) = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\mathsf{T}\right] = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_D) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_D) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_D, X_1) & \mathrm{cov}(X_D, X_2) & \cdots & \mathrm{var}(X_D) \end{bmatrix}.$$



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

Figure 1: Four pairs of random variables and their correlation coefficients.

1.2 The Gaussian Distribution

The density function of a univariate Gaussian (or normal) distribution with mean μ and variance σ^2 is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3}$$

We denote this distribution by $\mathcal{N}(\mu, \sigma^2)$ and denote $X \sim \mathcal{N}(\mu, \sigma^2)$ to denote that the random variable X is distributed according the normal distribution with mean μ and variance σ^2.

A multivariate Gaussian distribution in D dimensions has density function given by:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{4}$$

In this case, $\boldsymbol{\mu} \in \mathbb{R}^D$ is the mean and $\mathbf{\Sigma}$ is the covariance matrix.

1.3 The Laplace Distribution

The univariate Laplace distribution with parameters μ and b is defined by the density function.

$$\text{Lap}(x \mid \mu, b) = \frac{1}{2b} \cdot \exp\left(-\frac{|x-\mu|}{b}\right) \tag{5}$$

The mean of the $\text{Lap}(\cdot \mid \mu, b)$ is μ and the variance is $2b^2$.

1.4 Maximum Likelihood Principle

Suppose we are given some observations $x_1, x_2, \ldots, x_N$ drawn independently from some *unknown* distribution. For some distribution p, we define the likelihood of observing $x_1, \ldots, x_N$ as the probability of making these observations assuming that they had been generated according to p. Let us assume that p has some parametric form, for example, if p is a univariate Gaussian then p is entirely determined by the parameters μ and σ^2. Let θ denote the set of parameters determining p. Then, we can express the likelihood of observing the data $x_1, \ldots, x_N$ under the distribution p with parameters θ as follows:

$$p(x_1, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

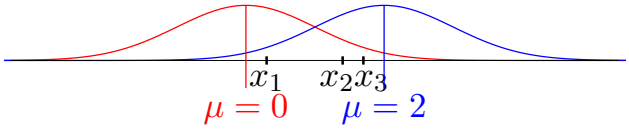


Figure 2: Maximum likelihood example

Since the observations $x_1, \ldots, x_N$ are independent, the joint distribution of the N observations is simply the product of the distribution of every observation. The maximum likelihood principle states that the parameters θ which have the highest likelihood should be picked.

To make this concrete, suppose you were given three points $x_1 = 0.3$, $x_2 = 1.4$ and $x_3 = 1.7$, with the promise that they were either generated according to the normal distribution $\mathcal{N}(0, 1)$ or $\mathcal{N}(2, 1)$. Which of the two would you say is more likely? We can write the likelihood as above and compare the two probabilities and respond with whichever is higher.

In this example, it is easy to see this just using a picture (see Fig 2). We see that point x_1 is as far from $\mu = 0$ as x_3 is from $\mu = 2$, and similarly x_3 is as far from $\mu = 0$ as x_1 is from $\mu = 2$. Thus, only x_2 will determine which of the two probabilities is more likely. In this case, since x_2 is closer to $\mu = 2$ than $\mu = 0$, we can conclude that $\mathcal{N}(2, 1)$ is more likely to generate these points. This only worked since both $\mathcal{N}(0, 1)$ and $\mathcal{N}(2, 1)$ had the same variance, otherwise we'd have had to do the calculations!

2 Linear Regression and Maximum Likelihood

Let us return to the linear model we studied in the previous lecture. We assumed that the output variable was a linear function of the input variables plus a noise term.

$$y = w_0 x_0 + w_1 x_1 + \cdots + w_D x_D + \epsilon \tag{6}$$

In the above expression, we've assumed that there is an extra input (or feature) x_0 which always takes the value 1, so that the constant term in the linear model does not have to be treated separately.

We model y (or alternatively the noise term ϵ) explicitly as a random variable. In particular, we'll model y as a random variable with mean $\mathbf{w} \cdot \mathbf{x}$. More formally,

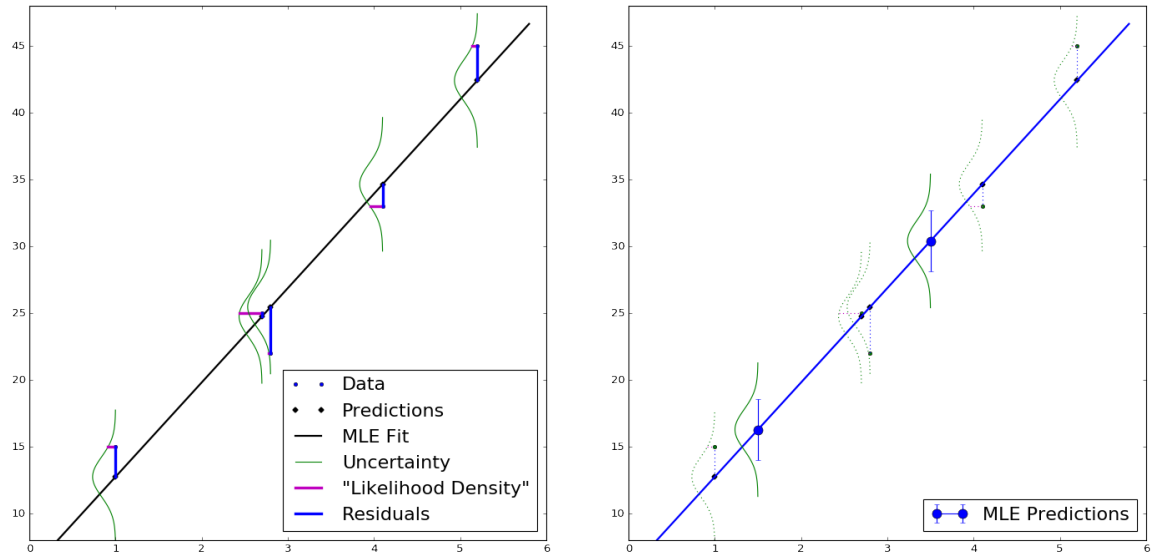
$$\mathbb{E}\left[y \mid \mathbf{w}, \mathbf{x}\right] = \mathbf{w}^\mathsf{T}\mathbf{x} \tag{7}$$

Thus, y given the inputs $\mathbf{x}$ and parameters $\mathbf{w}$ is modelled as a random variable with mean $\mathbf{w} \cdot \mathbf{x}$. In fact, we can further model y using a specific distribution. Let us model y conditioned on $\mathbf{x}$ and $\mathbf{w}$ as a Gaussian random variable with mean $\mathbf{w} \cdot \mathbf{x}$ and variance σ^2. Thus, in the language of probability, this linear model (with Gaussian noise) is expressed as

$$p(y \mid \mathbf{w}, \mathbf{x}) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}, \sigma^2) = \mathbf{w}^\mathsf{T}\mathbf{x} + \mathcal{N}(0, \sigma^2) \tag{8}$$

Alternatively, we can just think of modelling the noise term as a Gaussian random variable with mean 0 and variance σ^2.

Remark 1. *Throughout this lecture, we shall make no attempt to model the distribution over the inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Indeed, for simplicity we may as well think of them as fixed. The probabilistic model is only for the distribution of the output y given the input $\mathbf{x}$ and the model parameters $\mathbf{w}$. This is referred to as the* discriminative framework. *(When the inputs are also modelled as coming from a probability distribution, it is referred to as the* generative framework.*)*



(a) (b)

Figure 3: (a) Probabilistic version of the linear model with Gaussian noise. (b) Predictions using the maximum likelihood estimates.

Maximum Likelihood of Linear Regression with Gaussian Noise

Let us now turn to computing the likelihood of the observed data $\langle(\mathbf{x}_i, y_i)\rangle_{i=1}^N$ under the linear model with Gaussian noise. For every observation we have,

$$y_i = \mathbf{w} \cdot \mathbf{x}_i + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. We will make the assumption that the ϵ_i for $i = 1, \ldots, N$ are all independent random variables.

Before computing the likelihood mathematically, let us focus attention on Figure 3(a). The probabilistic interpretation of the linear model states that we actually expect the observations to deviate from the linear function (line) according to a normal distribution. In this example, the likelihood is given by the product of the lengths of the magenta segments (which are the values of the probability density function). The maximum likelihood estimate seeks to fit a line that maximises this product. On the other hand, the least squares estimate was minimising the sum of the squares of the residuals (the blue segments).

Now, let us write the likelihood of observing the data mathematically. Since, we only model the outputs $y_1, \ldots, y_N$ probabilisticially, we have

$$p\left(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma\right) = \prod_{i=1}^N p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, \sigma\right) \qquad \text{by independence}$$

According to the model $y_i \sim \mathbf{w}^\mathsf{T}\mathbf{x}_i + \mathcal{N}(0, \sigma^2)$

$$\begin{aligned} p\left(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma\right) &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^N (y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i)^2\right) \end{aligned}$$

We want to find parameters $\mathbf{w}$ and σ that maximise the likelihood. Since the logarithm, $\log : \mathbb{R}^+ \to \mathbb{R}$ is an increasing function, we can instead maximise the *log-likelihood* which can

be expressed in a slightly more convenient form.

$$\mathrm{LL}(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i)^2$$

We can express everything in vector notation,

$$\mathrm{LL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y})$$

Alternatively, we can minimise the *negative log-likelihood*

$$\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{2}\log(2\pi\sigma^2)$$

Recall the objective function we used for the least squares estimate in the previous lecture

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2N}(\mathbf{X}\mathbf{w} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y})$$

For minimizing with respect to $\mathbf{w}$, the two objectives $\mathcal{L}(\mathbf{w})$ and $\mathrm{NLL}(\mathbf{w})$ are the same upto a constant additive and multiplicative factor! Thus, we know that the maximum likelihood estimate for $\mathbf{w}$ is given by,

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

We can also find the maximum likelihood estimate for σ. An exercise on sheet 2 is to show that the MLE for σ is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}(\mathbf{X}\mathbf{w}_{\mathrm{ML}} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w}_{\mathrm{ML}} - \mathbf{y})$$

Prediction

Having obtained the maximum likelihood estimates (MLEs) $\mathbf{w}_{\mathrm{ML}}$ and σ^2_{ML}, on a new point $\mathbf{x}_{\mathrm{new}}$, we can use these to make a prediction and also give confidence intervals (see Figure 3(b)).

$$\widehat{y}_{\mathrm{new}} = \mathbf{w}_{\mathrm{ML}} \cdot \mathbf{x}_{\mathrm{new}}$$
$$p(y_{\mathrm{new}} \mid \mathbf{x}_{\mathrm{new}}, \mathbf{w}_{\mathrm{ML}}) = \widehat{y}_{\mathrm{new}} + \mathcal{N}(0, \sigma^2_{\mathrm{ML}})$$

Discussion

In this lecture, we viewed the linear model through a proabilistic framework, where the noise term is modelled explicitly according to a probability distribution. We made the choice to model this as a Gaussian random variable with mean 0 and variance σ^2; but of course, as we shall see shortly other choices are possible.

Once this model is expressed in the language of probability, we can apply the maximum likelihood principle which seeks to find parameters that maximise the likelihood of the observations under the chosen model. In order to find the MLE $\mathbf{w}_{\mathrm{ML}}$ for linear regression with Gaussian noise, the optimisation problem turned out to be exactly the same as that for obtaining the least squares estimate. Using this, we are also able to obtain the MLE for σ. Thus, in a way this can be seen as a justification of the least squares approach, although, it still remains to be justified that a Gaussian random variable is a suitable choice for modelling noise.

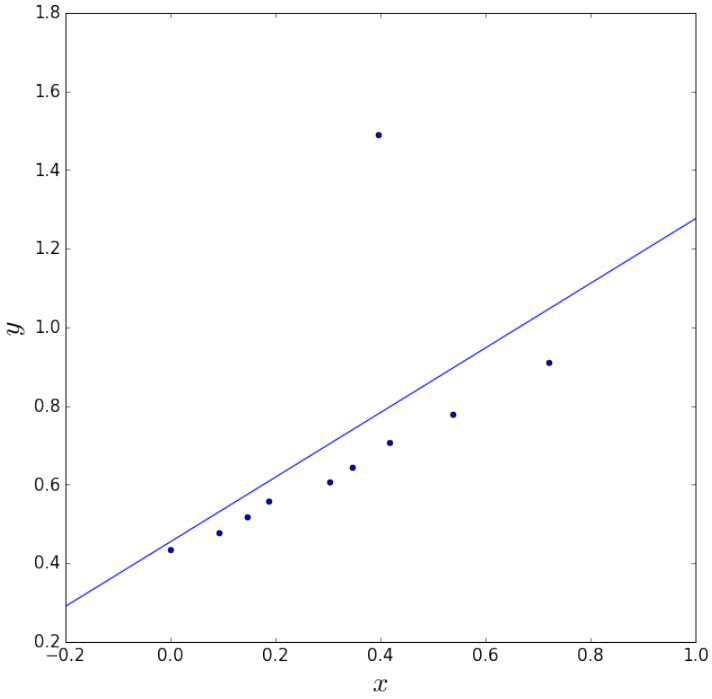


Figure 4: For linear regression, the least squares estimate and the MLE (with Gaussian noise) are not robust to outliers.

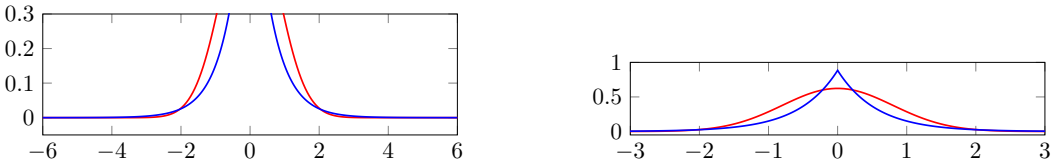


Figure 5: Laplace and Gaussian distribution with the same mean and variance.

2.1 Outliers and the Laplace Noise Model

Recall that in the previous lecture, we saw that the least squares estimate is very sensitive to outliers (see Fig. 4). As the maximum likelihood estimate under the Gaussian noise model results in the same estimate for the parameters $\mathbf{w}$, this too must be sensitive to noise! Probabilistically, we may view this as follows: the Gaussian distribution has very light 'tails', *i.e.,* there is very little probability mass a couple of standard deviations away from the mean. Thus under this model outliers are very very unlikely and so the model will not treat them as such and try to fit a model that accounts for them rather than ignoring them. Instead, we can model the noise using a distribution that has heavier tails. of which the Laplace distribution is one. Figure 5 shows the Laplace distribution and the Gaussian distribution with the same mean and variance. Although, it is a bit hard to see, from the zoomed in version it is clear that the tails of the Laplace distribution are heavier. Recall that the Laplace distribution with parameters μ and b is given by,

$$\text{Lap}(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

We can express the linear model with Laplace noise as:

$$p(y \mid \mathbf{w}, \mathbf{x}) = \text{Lap}(y \mid \mathbf{w} \cdot \mathbf{x}, b) = \mathbf{w} \cdot \mathbf{x} + \text{Lap}(\epsilon \mid 0, b)$$

As in the case of the Gaussian noise model, we can write the likelihood, log-likelihood and negative log-likelihood for the Laplace noise model.

$$\begin{aligned} p(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, b) &= \prod_{i=1}^{N} \frac{1}{2b} \exp\left(-\frac{|y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i|}{b}\right) \\ &= \frac{1}{(2b)^N} \exp\left(-\frac{1}{b}\sum_{i=1}^{N} |y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i|\right) \end{aligned}$$

As in the case of the Gaussian noise model, we look at the negative log-likelihood

$$\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, b) = \frac{1}{b}\sum_{i=1}^{N} |y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i| + N\log(2b)$$

We can see that the maximum likelihood estimate (in the Laplace noise model) is that which minimises the average absolute difference between the predictions and the observed outputs. This is exactly what we discussed in the previous lecture as a means to handle data with outliers. Recall, that there is no closed form expression for the solution to this optimisation problem. We shall study algorithms to solve this problem next week. Deriving the MLE for b is left as an exercise.

3 Information, Entropy, KL Divergence

We'll briefly discuss the connections between some of the concepts introduced in this lecture to those in information theory. Obviously, give that the goal of machine learning is to extract meaninful patterns out of data, it is no surprise that there are deep connections between machine learning and information theory. Exploring these in detail is beyond the scope of this course, but the interested student may refer to the book by MacKay (2003) or Jaynes (2003).

3.1 Entropy

Let X be a random variable that takes values from a finite set according to distribution p.[1] Then then entropy of X is defined as

$$H(X) = -\sum_{x} p(x)\log p(x) \tag{9}$$

The entropy is a measure of uncertainty of a random variable. If X takes values over a finite set of size n, then X has maximum entropy if it is distributed according to the uniform distribution over these n elements. It has minimum entropy if all the probability mass is concentrated on one of these elements, *i.e.,* in effect it is not a random variable at all, but a constant.

Let us focus on the case of Bernoulli random variables. A Bernoulli random variable is defined by a parameter $\theta \in [0, 1]$ and takes value 1 with probability θ and 0 with probability $1 - \theta$. This can be expressed succintly as

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$$

In this case, let us write the entropy in terms of the parameter θ and use logarithm base 2 for convenience.

$$H(X) = -\theta\log_2(\theta) - (1-\theta)\log_2(1-\theta)$$

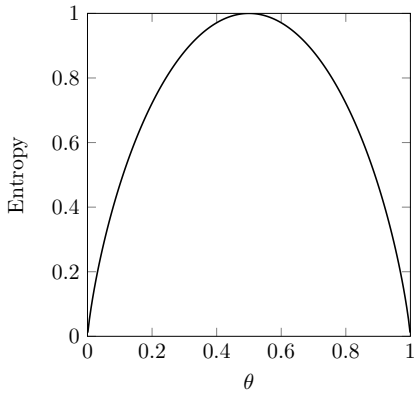


Figure 6: Entropy of the Bernoulli random variable as a function of θ

Figure 6 plots the entropy as a function of θ. We see that the entropy has a maximum value of 1 for $\theta = 1/2$ and minimum value of 0 at $\theta \in \{0, 1\}$. One way to think of entropy is how much information is obtained when the outcome of an experiment is revealed. For example, if Alice has an unbiased coin, then if she tosses it and reports the outcome we get one *bit* of information. On the other hand if she has a coin that always lands on heads, we get no additional information by being told the outcome of the coin toss, because it was something we could have predicted ourselves with complete certainty!

3.2 Kullback-Leibler Divergence

Let p and q be distributions over some finite set and suppose that the support of p is contained in the suport of q.[1] The Kullback-Leibler (or KL) Divergence between two distributions p and q is defined as follows

$$\begin{aligned}\mathrm{KL}(p\|q) &= \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) \\ &= \sum_x p(x)\log(p(x)) - \sum_x p(x)\log(q(x)) = -H(p) + H(p,q)\end{aligned}$$

Here $H(p) = -\sum_x p(x)\log p(x)$ is the entropy of the distribution p and $H(p,q) = -\sum_x p(x)\log q(x)$ is called the *cross-entropy*. The cross entropy accounts of the expected number of bits required to encode an observation from p if the encoding scheme was based on q. Thus, the KL-divergence $\mathrm{KL}(p\|q)$ gives the expected *excess bits* required to encode an observation from p if the encoding scheme was based on q.

The KL divergence satisfies the following two properties:

1. $\mathrm{KL}(p\|q) \geq 0$
2. $\mathrm{KL}(p\|q) = 0$ if and only if $p = q$

It is worth mentioning that the KL-divergence is not a distance; in particular, it is not symmetric. For example, even when the support of p and q is the same, so that both $\mathrm{KL}(p\|q)$ and $\mathrm{KL}(q\|p)$ are defined, they need not be equal.

[1] This can be extended to continuous-valued random variables by using the integral instead of the sum and replacing the probability mass function by the density function.

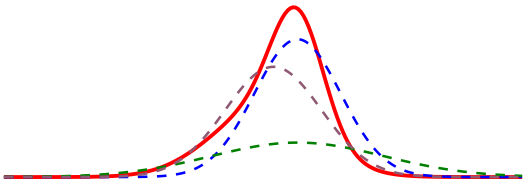


Figure 7: Maximum Likelihood Estimates and KL-divergence: The data is generated according to the distribution shown by the thick red line. The figure also shows three possible Gaussian distributions (dashed). The goal is to find the Gaussian distribution that maximises the likelihood of the observed data.

Relation to Maximum Likelihood

Let us now see how the maximum likelihood estimate relates to these notions from information theory. Suppose we get data $x_1, \dots x_N$ from some unknown distribution p (not necessarily of any particular parametric form). However, we wish to fit a distribution that does have some parametric form (say for example Gaussian) that best explains the data. In particular, we will derive the maximum likelihood estimate for the parameters of distributions of a certain parametric form.

Figure 7 shows the actual generating distribution (in thick red). It also shows three possible Gaussian distributions with different means and variances (dotted). Suppose, we want to find maximum likelihood estimate for these parameters.

The mathematical derivation below is more general. It just assumes that the family of distributions we consider are parameterized by some parameters θ. In particular, $q(\cdot \mid \theta)$ is the distribution that we use to model the data and we derive the maximum likelihood estimate for θ.

$$
\begin{aligned}
\widehat{\theta}_{\mathrm{ML}} &= \operatorname*{argmax}_{\theta} \prod_{i=1}^{N} q(x_i \mid \theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{i=1}^{N} \log(q(x_i \mid \theta)) \\
&= \operatorname*{argmax}_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log(q(x_i \mid \theta)) - \frac{1}{N} \sum_{i=1}^{N} \log(p(x_i)) && (10) \\
&= \operatorname*{argmin}_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{p(x_i)}{q(x_i \mid \theta)} \right) && (11) \\
&\underset{N \to \infty}{\longrightarrow} \operatorname*{argmin}_{\theta} \int \log \left(\frac{p(x)}{q(x|\theta)} \right) p(x) dx = \mathrm{KL}(p \| q_\theta) && (12)
\end{aligned}
$$

Above in Step (10) we replace the sum by the average and added an extra term that does not depend on θ, neither of these operations affects the argmax; in Step (11), we switched the signs and hence changed the argmax to argmin; finally, Step (12) states that in the limit of getting infinite quantities of data, where $x_i \sim p$, the average can be replaced by the expectation under the distribution p. This last term is nothing but the KL-divergence $\mathrm{KL}(p \| q_\theta)$. Thus, the maximum likelihood estimate can be viewed as finding parameters (from some family of distributions) that minimises the KL-divergence between the true distribution generating the data and the modelled distribution from this family. Alternatively, the MLE can be viewed

as finding the distribution from a parametric family that has least KL-divergence between the empirical distribution over the data and this particular parametric distribution.

Remark 2. *This section covered somewhat advanced topics and is not examinable. It is introduced to show connections between machine learning methods and information theory.*

References

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Kevin P. Murphy. *Machine Learning : A Probabilistic Perspective*. MIT Press, 2012.