Chatper 1: Introduction

ifding

Abstract

Probability Theory, Decision Theory, Information Theory

1. Probability Theory

1.1. The Rules of Probability

sum rule:

$$p(X) = \sum_{Y} p(X, Y) \tag{1}$$

product rule:

$$p(X, Y) = p(Y|X)p(X) \tag{2}$$

Here $p(X, Y)$ is a joint probability and is verbalized as "the probability of X and Y". $p(Y|X)$ is a conditional probability and is verbalized as "the probability of Y given X". $p(X)$ is a marginal probability and is simply "the probability of X". These two simple rules form the basis for all of the probabilistic machinery.

From the symmetry property $p(X, Y) = p(Y, X)$, we can obtain *Bayes' theorem*:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{3}$$

Using the sum rule, the denominator in Bayes' theorem can be expressed in:

$$p(X) = \sum_{Y} p(X|Y)p(Y) \tag{4}$$

We can view the denominator in Bayes' theorem as being the normalization constant required to ensure that the sum of the conditional probability $p(Y|X)$ over all values of Y equals 1.

1.2. Probability densities

Considering probabilities defined over continuous variables, If the probability of a real-valued variable x falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$, then $p(x)$ is called the *probability density* over x.

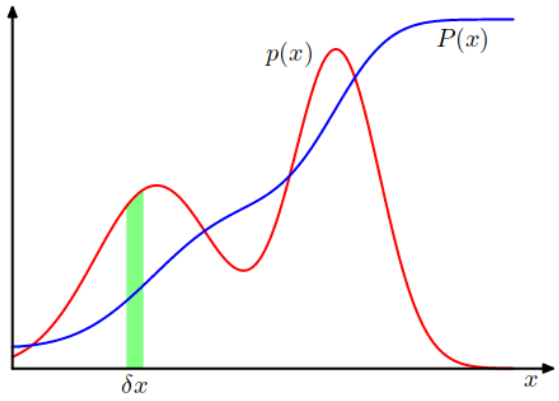

Figure 1: The probability density

$$p(x \in (a, b)) = \int_{a}^{b} p(x)\mathrm{d}x \tag{5}$$

The probability density $p(x)$ must satisfy the two conditions:

$$\begin{aligned} p(x) &\geqslant 0 \\ \int_{-\infty}^{\infty} p(x)\mathrm{d}x &= 1 \end{aligned} \tag{6}$$

The probability that x lies in the interval $(-\infty, z)$ is given by the *cumulative distribution function* defined by

$$P(z) = \int_{-\infty}^{z} p(x)\mathrm{d}x \tag{7}$$

Note that if x is a discrete variable, $p(x)$ is called a *probability mass function* because it can be regarded as

a set of 'probability masses' concentrated at the allowed values of x.

The sum and product rules of probability, as well as Bayes' theorem, apply equally to the case of probability densities, or to combinations of discrete and continuous variables. For instance, if x and y are two real variables,

$$p(x) = \int p(x, y) \mathrm{d}y \tag{8}$$

$$p(x, y) = p(y|x)p(x) \tag{9}$$

1.3. Expectations and covariances

One of the most important operations involving probabilities is that of finding **weighted averages** of functions. The average value of function $f(x)$ under a probability distribution $p(x)$ is called the *expectation* of $f(x)$. The average is weighted by the relative probabilities of the different values of x.

$$\mathbb{E}[f] = \sum_x p(x) f(x) \tag{10}$$

In the case of continuous variables, expectations are expressed by the corresponding probability density.

$$\mathbb{E}[f] = \int p(x) f(x) \mathrm{d}x \tag{11}$$

If we are given a finite number N of points drawn from the probability distribution or probability density, then the expectation can be approximated as a finite sum over these points

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f\left(x_n\right) \tag{12}$$

Sometimes we will be considering expectations of functions of several variables, $\mathbb{E}_x[f(x, y)]$ denotes the average of the function $f(x, y)$ with respect to the distribution of x. It will be a function of y. A conditional expectation with respect to a conditional distribution

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x) \tag{13}$$

The *variance* of $f(x)$ is defined by

$$\begin{aligned} \mathrm{var}[f] &= \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \\ &= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[f(x)]^2 \end{aligned} \tag{14}$$

It provides a measure of how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$.

In particular, we can consider the variance of the variable x itself,

$$\mathrm{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2 \tag{15}$$

For two random variables x and y, the *covariance* is defined by

$$\begin{aligned} \mathrm{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \tag{16}$$

which expresses the extent to which x and y vary together. If x and y are independent, then their covariance vanishes.

In the case of two vectors of random variables $\mathbf{x}$ and $\mathbf{y}$, the covariance is a matrix

$$\begin{aligned} \mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\left\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}\left[\mathbf{y}^{\mathrm{T}}\right]\right\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\mathbf{x}\mathbf{y}^{\mathrm{T}}\right] - \mathbb{E}[\mathbf{x}]\mathbb{E}\left[\mathbf{y}^{\mathrm{T}}\right] \end{aligned} \tag{17}$$

1.4. Bayesian probabilities

The more general Bayesian view: probabilities provide a quantification of uncertainty.

We can describe the uncertainty in model parameters $\mathbf{w}$, we capture our assumptions about $\mathbf{w}$, before observing the data, in the form of a prior probability distribution $p(\mathbf{w})$.

The effect of the observed data $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ is expressed through the conditional probability $p(\mathcal{D}|\mathbf{w})$.

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{18}$$

then allows us to evaluate the uncertainty in $\mathbf{w}$ *after* we have observed $\mathcal{D}$ in the form of the posterior probability $p(\mathbf{w}|\mathcal{D})$.

The quantity $p(\mathcal{D}|\mathbf{w})$ is evaluated for the observed data set $\mathcal{D}$ and can be viewed as a function of the parameter vector $\mathbf{w}$, it is called the *likelihood function*. It expresses

how probable the observed data set is for different settings of the parameter vector $\mathbf{w}$.

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{19}$$

where all of these quantities are viewed as functions of $\mathbf{w}$. $p(\mathcal{D})$ is the normalization constant, which ensures that the posterior distribution is a valid probability density and integrates to one.

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w} \tag{20}$$

maximum likelihood: $\mathbf{w}$ is set to the value that maximizes the likelihood function $p(\mathcal{D}|\mathbf{w})$. This corresponds to choosing the value of $\mathbf{w}$ for which the probability of observed data set is maximized. The negative log of the likelihood function is called an *error function*. Maximizing the likelihood is equivalent to minimizing the error.

1.5. The Gaussian distribution

For the case of a single real-valued variable x, the Gaussian distribution is defined by

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{21}$$

which is governed by two parameters: μ, called the *mean*, and σ^2, called the *variance*. σ is called the *standard deviation* and $\beta = 1/\sigma^2$ is called the *precision*.

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\mathrm{d}x = 1 \tag{22}$$

The average value of x is given by

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x\mathrm{d}x = \mu \tag{23}$$

For the second order moment

$$\mathbb{E}\left[x^2\right] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2\mathrm{d}x = \mu^2 + \sigma^2 \tag{24}$$

It follows that the variance of x is given by

$$\mathrm{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2 = \sigma^2 \tag{25}$$

Gaussian distribution over a D-dimensional vector $\mathbf{x}$ of continuous variables:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{26}$$

where the D-dimensional vector μ is called the mean, the $D \times D$ matrix $\boldsymbol{\Sigma}$ is called the covariance, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

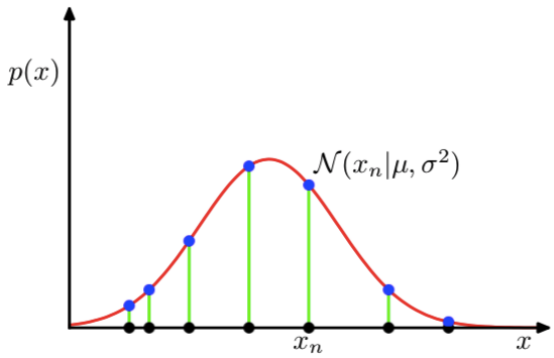

Figure 2: The likelihood function ($p(\mathcal{D}|\mathbf{w})$) for a Gaussian distribution, shown by the red curve. Here the black points denote a data set of values $\{x_n\}$.

A data set of observations $\mathbf{X} = (x_1,\ldots,x_N)^{\mathrm{T}}$, representing N observations of the scalar variable x. $\mathbf{x}$ denotes a single observation of the vector-valued variable $(x_1,\ldots,x_D)^{\mathrm{T}}$. Our data set $\mathbf{X}$ is *independent and identically distributed*, which is often abbreviated to i.i.d.

$$p\left(\mathbf{X}|\mu,\sigma^2\right) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right) \tag{27}$$

As shown in Figure 2, the likelihood function given by the above equation corresponds to the product of the blue values. Maximizing the likelihood involves adjusting the mean μ and variance σ^2 of the Gaussian so as to maximize this product. The log likelihood function can be written in the form

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{28}$$

We obtain the maximum likelihood solution given by

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{29}$$

which is the sample mean i.e., the mean of the observed values $\{x_n\}$.

We obtain the maximum likelihood solution for the variance in the form

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 \tag{30}$$

The **significant limitations** of the maximum likelihood approach: the maximum likelihood approach systematically underestimates the variance of the distribution. A phenomenon called *bias* and *overfitting*.

The maximum likelihood solutions μ_{ML} and σ^2_{ML} are functions of the data set values $x_1, \ldots, x_N$. Consider the expectations of these quantities,

$$\begin{aligned}\mathbb{E}\left[\mu_{\mathrm{ML}}\right] &= \mu \\ \mathbb{E}\left[\sigma^2_{\mathrm{ML}}\right] &= \left(\frac{N-1}{N}\right)\sigma^2\end{aligned} \tag{31}$$

so that on average the maximum likelihood estimate will obtain the correct mean but will underestimate the true variance by a factor $(N-1)/N$.

As shown in Figure 3, the green curve shows the true Gaussian distribution from which data is generated, the three red curves show the Gaussian distributions obtained by fitting to three data sets, each consisting of two data points shown in blue. Averaged across the three data sets, the mean is correct, but the variance is systematically under-estimated because it is measured relative to the sample mean and not relative to the true mean.

Note that the bias of the maximum likelihood solution becomes less significant as the number N of data points increases. For the more complex models, the bias problems associated with maximum likelihood will be much more severe. The issue of bias in maximum likelihood lies at the root of the over-fitting problem.

2. Decision Theory

Suppose we have an input vector $\mathbf{x}$ together with a corresponding vector $\mathbf{t}$ of target variables, and our goal is to predict $\mathbf{t}$ given a new value for $\mathbf{x}$. For regression problems, $\mathbf{t}$ will comprise continuous variables. For classification problems, $\mathbf{t}$ will represent class labels.

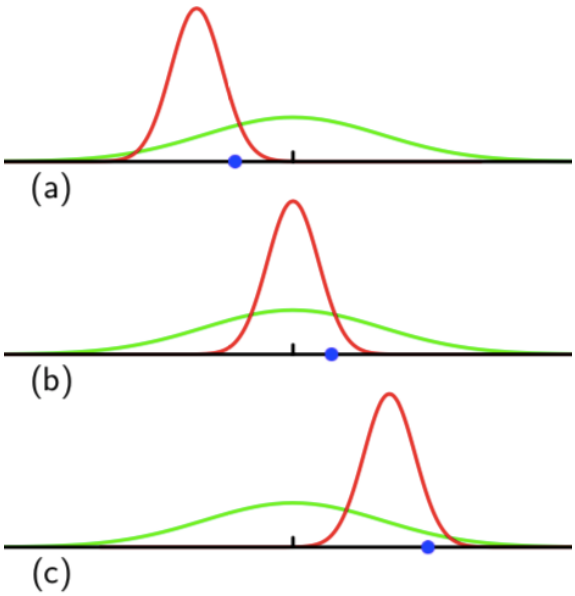

Figure 3: Illustration of how bias arises in using maximum likelihood to determine the variance of a Gaussian.

The joint probability distribution $p(\mathbf{x}, \mathbf{t})$ provides a complete summary of the uncertainty associated with these variables. Determination of $p(\mathbf{x}, \mathbf{t})$ from a set of training data is an example of *inference*.

Consider, for example, when we obtain the X-ray image $\mathbf{x}$, our goal is to decide which of the two classes to assign to the image. We are interested in the probabilities of the two classes given the image $p(\mathcal{C}_k|\mathbf{x})$.

$$p\left(\mathcal{C}_k|\mathbf{x}\right) = \frac{p\left(\mathbf{x}|\mathcal{C}_k\right)p\left(\mathcal{C}_k\right)}{p(\mathbf{x})} \tag{32}$$

Here, $p(\mathcal{C}_k)$ as the prior probability for the class $\mathcal{C}_k$, and $p(\mathcal{C}_k|\mathbf{x})$ as the corresponding posterior probability. If our aim is to minimize the chance of assigning $\mathbf{x}$ to the wrong class, then intuitively we would choose the class having the higher posterior probability.

We need to divide the input space into *decision regions* $\mathcal{R}_k$, one for each class, such that all points in $\mathcal{R}_k$ are assigned to class $\mathcal{C}_k$.

$$\begin{aligned}p(\text{ mistake }) &= p\left(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2\right) + p\left(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1\right) \\ &= \int_{\mathcal{R}_1} p\left(\mathbf{x}, \mathcal{C}_2\right)\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p\left(\mathbf{x}, \mathcal{C}_1\right)\mathrm{d}\mathbf{x}\end{aligned} \tag{33}$$

If $p\left(\mathbf{x}, \mathcal{C}_1\right) > p\left(\mathbf{x}, \mathcal{C}_2\right)$ for a given value of $\mathbf{x}$, then we should assign that $\mathbf{x}$ to class $\mathcal{C}_1$. Because $p\left(\mathbf{x}, \mathcal{C}_k\right) = p\left(\mathcal{C}_k|\mathbf{x}\right)p(\mathbf{x})$, the factor $p(\mathbf{x}$ is common to both terms.

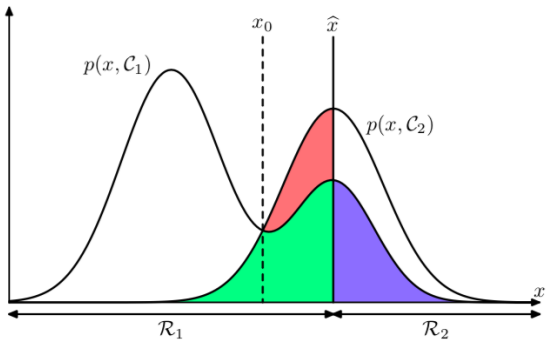

Figure 4: The joint probabilities $p(\mathbf{x}, \mathcal{C}_k)$ for each of two classes, with the decision boundary $x = \hat{x}$, Errors arise from the blue, green, and red regions.

In Figure 4, values of $x \geq \hat{x}$ are classified as $\mathcal{C}_2$, whereas they are classified as $\mathcal{C}_1$. For $x < \hat{x}$, the errors are due to points from class $\mathcal{C}_2$ being misclassified as $\mathcal{C}_1$ (the red and green regions).In the region $x \geq \hat{x}$, the errors are due to points from class $\mathcal{C}_1$ being misclassified as $\mathcal{C}_2$ (the blue region). As changing the location of $\hat{x}$, the combined areas of the green and blue regions remains constant, whereas the size of the red region varies. When set $\hat{x} = x_0$, the red region disappears, this is minimum misclassification rate decision rule, which assigns each value of x to the class having the higher posterior probability $p(\mathcal{C}_k|x)$.

3. Information Theory

The amount of information can be viewed as the 'degree of surprise' on learning the value of random variable x. If a highly improbable event has just occurred, we will have received more information than some very likely event has just occurred.

$$h(x) = -\log_2 p(x) \tag{34}$$

where the negative sign ensures that information is positive or zero. Note that low probability events x correspond to high information content. The *entropy* of the discrete random variable x:

$$\mathrm{H}[x] = -\sum_x p(x) \ln p(x) \tag{35}$$

Note that $\lim_{p\to 0} p \ln p = 0$ and so $p(x) \ln p(x) = 0$. For continuous variables, the differential entropy is given by

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{36}$$

Suppose we have a joint distribution $p(\mathbf{x}, \mathbf{y})$ from which we draw pairs of values of $\mathbf{x}$ and $\mathbf{y}$. If a value of $\mathbf{x}$ is already known, then the additional information needed to specify the corresponding value of $\mathbf{y}$ is given by $\ln p(\mathbf{y}|\mathbf{x})$. Thus the average additional information needed to specify $\mathbf{y}$ (*conditional entropy*) can be written as

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \tag{37}$$

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}] \tag{38}$$

where $\mathrm{H}[\mathbf{x}, \mathbf{y}]$ is the differential entropy of $p(x, y)$ and $H[\mathbf{x}]$ is the differential entropy of the marginal distribution $p(\mathbf{x})$.

3.1. KL divergence

Consider some unknown distribution $p(\mathbf{x})$, and suppose that we have modelled this using an approximating distribution $q(\mathbf{x})$.

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x} \right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x} \end{aligned} \tag{39}$$

This is known as the *relative entropy* or *KL divergence*, between the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$. Note that it is not a symmetrical quantity, $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$.

$\mathrm{KL}(p||q) \geq 0$, if and only if, $p(\mathbf{x}) = q(\mathbf{x})$. To prove it, we first introduce the concept of *convex* functions.

$$f(\lambda a + (1-\lambda)b) \leqslant \lambda f(a) + (1-\lambda) f(b) \tag{40}$$

A convex function $f(x)$ satisfies:

$$f\left(\sum_{i=1}^{M} \lambda_i x_i \right) \leqslant \sum_{i=1}^{M} \lambda_i f\left(x_i\right) \tag{41}$$

This result is known as *Jensen's inequality*. If we interpret the λ_i as the probability distribution over a discrete variable x, then

$$f(\mathbb{E}[x]) \leqslant \mathbb{E}[f(x)] \tag{42}$$

where $\mathbb{E}[\cdot]$ denotes the expectation. For continuous variables, Jensen's inequality takes the form

$$f\left(\int \mathbf{x}p(\mathbf{x})\mathrm{d}\mathbf{x}\right) \leqslant \int f(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{43}$$

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x} \geqslant -\ln\int q(\mathbf{x})\mathrm{d}\mathbf{x} = 0 \tag{44}$$

We can interpret the KL divergence as a measure of the distance of the two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$.

Suppose that data is being generated from an unknown distribution $p(\mathbf{x})$ that we wish to model. We can try to approximate this distribution using some parametric distribution $q(\mathbf{x}|\theta)$, governed by a set of adjustable parameters θ, for example a multivariate Gaussian.

One way to determine θ is to minimize the KL divergence between $p(\mathbf{x})$ and $q(\mathbf{x}|\theta)$ with respect to θ. We cannot do this directly because we don't know $p(\mathbf{x})$. Suppose, however, that we have observed a finite set of data points $\mathbf{x}_n$, for $n = 1, \ldots, N$, drawn from $p(\mathbf{x})$.

Then the expectation with respect to $p(\mathbf{x})$ can be approximated by a finite sum over these points, using Eq. 12, so that

$$\mathrm{KL}(p\|q) \simeq \sum_{n=1}^{N} \left\{-\ln q\left(\mathbf{x}_n|\boldsymbol{\theta}\right) + \ln p\left(\mathbf{x}_n\right)\right\} \tag{45}$$

The second term on the right-hand side is independent of θ, and the first term is the negative log likelihood function for θ under the distribution $q(\mathbf{x}|\theta)$ evaluated using the data set. Thus we see that minimizing this KL divergence is equivalent to maximizing the likelihood function.

3.2. mutual information

Now consider the joint distribution between two sets of variables $\mathbf{x}$ and $\mathbf{y}$ given by $p(\mathbf{x}, \mathbf{y})$. If they are independent, $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. If the variables are not independent, we can gain some idea of whether whey are 'close' to being independent by

$$\begin{aligned}\mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})}\right)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\end{aligned} \tag{46}$$

which is called the *mutual information* between the variables x and y. $\mathrm{I}[\mathbf{x}, \mathbf{y}] \geq 0$ with equality if, and only if, x and y are independent.

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}] \tag{47}$$

Thus we can view the mutual information as the reduction in the uncertainty about $\mathbf{x}$ by virtue of being told the value of $\mathbf{y}$ (or vice versa).

From a Bayesian perspective, we can view $p(\mathbf{x})$ as the prior distribution for x and $p(\mathbf{x}|\mathbf{y})$ as the posterior distribution after we have observed new data $\mathbf{y}$. The mutual information therefore represents the reduction in uncertainty about $\mathbf{x}$ as a consequence of the new observation $\mathbf{y}$.

4. Exercises

1.6 Based on the following equation:

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\left\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}\left[\mathbf{y}^{\mathrm{T}}\right]\right\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\mathbf{x}\mathbf{y}^{\mathrm{T}}\right] - \mathbb{E}[\mathbf{x}]\mathbb{E}\left[\mathbf{y}^{\mathrm{T}}\right]\end{aligned} \tag{48}$$

Because $\mathbf{x}$ and $\mathbf{y}$ is independent, we have: $p(\mathbf{x}, \mathbf{y}) = p_{\mathbf{x}}(\mathbf{x})p_{\mathbf{y}}(\mathbf{y})$, therefore:

$$\begin{aligned}\iint xyp(x, y)dxdy &= \iint xyp_x(x)p_y(y)dxdy \\ &= \left(\int xp_x(x)dx\right)\int yp_y(y)dy \\ \Rightarrow \mathbb{E}_{x,y}[xy] = \mathbb{E}[x][\mathrm{E}[y] &\Rightarrow \mathrm{cov}[\mathbf{x}, \mathbf{y}] = 0\end{aligned} \tag{49}$$

1.10 Suppose that two variables x and z are statistically independent.

$$\begin{aligned}\mathbb{E}[x + z] &= \iint (x + z)p(x, z)dxdz \\ &= \iint (x + z)p(x)p(z)dxdz \\ &= \iint xp(x)p(z)dxdz + \iint zp(x)p(z)dxdz \\ &= \int\left(\int p(z)dz\right)xp(x)dx + \int\left(\int p(x)dx\right)zp(z)dz \\ &= \int xp(x)dx + \int zp(z)dz \\ &= \mathbb{E}(x] + \mathbb{E}[z]\end{aligned} \tag{50}$$

$$
\begin{aligned}
\mathrm{var}[x+z] &= \iint (x+z-\mathbb{E}[x+z])^2 p(x,z)dxdz \\
&= \iint \{(x+z)^2 - 2(x+z)\mathbb{E}[x+z] + \\
&\mathbb{E}^2[x+z]\}p(x,z)dxdz \\
&= \iint (x+z)^2 p(x,z)dxdz - \\
&2\mathbb{E}[x+z]\iint (x+z)p(x,z)dxdz + \mathbb{E}^2[x+z] \\
&= \iint \left(x^2+z\right)^2 p(x,z)dxdz - \mathbb{E}^2[x+z] \\
&= \iint \left(x^2+2xz+z^2\right) p(x)p(z)dxdz - \mathbb{E}^2[x+z] \\
&= \mathbb{E}\left[x^2\right] + \mathbb{E}\left[z^2\right] - \mathbb{E}[z] - 2\mathbb{E}[x]\mathbb{E}[z] + \\
&2\iint xzp(x)p(z)dxdz \\
&= \mathrm{var}[x] + \mathrm{var}[z] - 2\mathbb{E}[x]\mathbb{E}[z] + \\
&2\left(\int xp(x)dx\right)\iint xzp(x)p(z)dxdz \\
&= \mathrm{var}[x] + \mathrm{var}[z]
\end{aligned} \tag{51}
$$

1.30 Evaluate the KL divergence between two Gaussians $p(x) = \mathcal{N}\left(x|\mu,\sigma^2\right)$ and $q(x) = \mathcal{N}\left(x|m,s^2\right)$.

Based on definition:

$$
\begin{aligned}
\ln\{\frac{p(x)}{q(x)}\} &= \ln\left(\frac{s}{\sigma}\right) - [\frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2s^2}(x-m)^2] \\
&= \ln\left(\frac{s}{\sigma}\right) - [\left(\frac{1}{2\sigma^2} - \frac{1}{2s^2}\right)x^2 \\
&- \left(\frac{\mu}{\sigma^2} - \frac{m}{s^2}\right)x + \left(\frac{\mu^2}{2\sigma^2} - \frac{m^2}{2s^2}\right)]
\end{aligned} \tag{52}
$$

We will take advantage of the following equations to solve this problem

$$
\begin{aligned}
\mathbb{E}\left[x^2\right] &= \int x^2 \mathcal{N}\left(x|\mu,\sigma^2\right)dx = \mu^2 + \sigma^2 \\
\mathbb{E}[x] &= \int x\mathcal{N}\left(x|\mu,\sigma^2\right)dx = \mu \\
&\int \mathcal{N}\left(x|\mu,\sigma^2\right)dx = 1
\end{aligned} \tag{53}
$$

$$
\begin{aligned}
KL(p\|q) &= -\int p(x)\ln\{\frac{q(x)}{p(x)}\}dx \\
&= \ln\left(\frac{s}{\sigma}\right) - \left(\frac{1}{2\sigma^2} - \frac{1}{2s^2}\right)(\mu^2+\sigma^2) + \\
&(\frac{\mu}{\sigma^2} - \frac{m}{s^2})\mu - (\frac{\mu^2}{2\sigma^2} - \frac{m^2}{2s^2}) \\
&= \ln\left(\frac{s}{\sigma}\right) + \frac{\sigma^2+(\mu-m)^2}{2s^2} - \frac{1}{2}
\end{aligned} \tag{54}
$$

1.31 Based on $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$, $\int p(\mathbf{x},\mathbf{y})d\mathbf{y} = p(\mathbf{x})$, we first calculate $H[\mathbf{x},\mathbf{y}]$:

$$
\begin{aligned}
H[\mathbf{x},\mathbf{y}] &= -\iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= -\iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x})d\mathbf{x}d\mathbf{y} \\
&- \iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y} \\
&= -\int p(\mathbf{x})\ln p(\mathbf{x})d\mathbf{x} - \iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y} \\
&= H[\mathbf{x}] + H[\mathbf{y}|\mathbf{x}]
\end{aligned} \tag{55}
$$

$$
H[\mathbf{x}] + H[\mathbf{y}] - H[\mathbf{x},\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]
$$

$$
\begin{aligned}
&= -\int p(\mathbf{y})\ln p(\mathbf{y})d\mathbf{y} + \iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y} \\
&= -\iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y})d\mathbf{x}d\mathbf{y} + \iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y} \\
&= -\iint p(\mathbf{x},\mathbf{y})\ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right)d\mathbf{x}d\mathbf{y} \\
&= KL(p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) = I(\mathbf{x},\mathbf{y}) \geq 0
\end{aligned} \tag{56}
$$

The mutual information

$$
I(\mathbf{x},\mathbf{y}) = -\iint p(\mathbf{x},\mathbf{y})\ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right)d\mathbf{x}d\mathbf{y} \tag{57}
$$

References

[1] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

[2] https://github.com/zhengqigao/PRML-Solution-Manual