Chatper 2: Probability Distributions

ifding

Abstract

Binary Variables, Multinomial Variables, Gaussian Distribution, Exponential Family, Nonparametric Methods

density estimation: model the probability distribution $p(x)$ of a random variable x, given a finite set $x_1, \ldots, x_N$ of observations. It should be emphasized that the problem of density estimation is **fundamentally ill-posed**, because there infinitely probability distributions that could have given rise to the observed finite data set. Indeed, any distribution $p(x)$ that is nonzero at each of the data points $x_1, \ldots, x_N$ is a potential candidate.

parametric distributions, e.g., binomial and multinomial distributions for discrete random variables and Gaussian distribution for continuous random variables, are governed by a small number of adaptive parameters, such as mean and variance in a Gaussian.

nonparametric density estimation methods: the distribution typically depends on the size of the data set. Such models still contain parameters, but these control the model complexity rather than the form of the distribution.

1. Binary Variables

Considering a binary random variable $x \in \{0, 1\}$ The probability of $x = 1$ will be denoted by the parameter μ, so that $p(x = 1|\mu) = \mu$, where $0 \leq \mu \leq 1$, $p(x = 0|\mu) = 1 - \mu$. The *Bernoulli* distribution:

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} \tag{1}$$

$$\begin{aligned} \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu) \end{aligned} \tag{2}$$

Suppose we have a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$, which are drawn independently from $p(x|\mu)$, the likelihood function:

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p\left(x_n|\mu\right) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \tag{3}$$

$$\begin{aligned} \ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p\left(x_n|\mu\right) = \\ \sum_{n=1}^{N} \left\{x_n \ln \mu + (1-x_n)\ln(1-\mu)\right\} \end{aligned} \tag{4}$$

The log likelihood function depends on the N observations x_n only through their sum $\sum_n x_n$, which provides an example of a *sufficient statistic* for the data under this distribution. If set the derivative of $\ln p(\mathcal{D}|\mu)$ with respect to μ equal to zero,

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{5}$$

which is known as the *sample mean*. If we denote the number of observations of $x = 1$ (heads) within this data set by m, then $\mu_{\mathrm{ML}} = m/N$. Suppose we flip a coin 3 times and happen to observe 3 heads. Then $N = m = 3$ and $\mu_{\mathrm{ML}} = 1$. In this case, the maximum likelihood result would predict that all future observations should give heads. This is an extreme example of the over-fitting associated with maximum likelihood.

The *binomial* distribution. To obtain the normalization coefficient, we have to add up all of the possible ways of obtaining m heads.

$$\begin{aligned} \mathrm{Bin}(m|N,\mu) &= \binom{N}{m} \mu^m (1-\mu)^{N-m} \\ \binom{N}{m} &\equiv \frac{N!}{(N-m)!m!} \end{aligned} \tag{6}$$

2. Multinomial Variables

1-of-K scheme: K-dimensional vector $\mathbf{x}$ in which one of the elements x_k equals 1, and all remaining elements equal 0. For example, $K = 6$ and $x_3 = 1$, then

$$\mathbf{x} = (0, 0, 1, 0, 0)^T. \tag{7}$$

Note that such vectors satisfy $\sum_{k=1}^{K} x_k = 1$. If we denote the probability of $x_k = 1$ by the parameter μ_k, then the distribution of $\mathbf{x}$ is given

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{8}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^T$, and the parameters μ_k are constrained to satisfy $\mu_k \geq 0$ and $\sum_k \mu_k = 1$, because they represent probabilities. The distribution $p(\mathbf{x}|\boldsymbol{\mu})$ can be regarded as a generalization of Bernoulli distribution to more than two outcomes. It's easily seen that the distribution is normalized

$$\begin{aligned} \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) &= \sum_{k=1}^{K} \mu_k = 1 \\ \mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} &= (\mu_1, \ldots, \mu_M)^{\mathrm{T}} = \boldsymbol{\mu} \end{aligned} \tag{9}$$

3. The Gaussian Distribution

It's widely used model for the distribution of continuous variables. In the case of a single variable x,

$$\mathcal{N}\left(x|\mu, \sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{10}$$

where μ is the mean and σ^2 is the variance. For a D-dimensional vector $\boldsymbol{x}$, the multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) =$

$$\frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\} \tag{11}$$

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

The *central limit theorem*, considering N variables x_1, ..., x_N each of which has a uniform distribution over the interval [0,1] and then considering the distribution of the mean $(x_1 + \ldots + x_N)/N$. For large N, this distribution tends to a Gaussian.

The functional dependence of the Gaussian on $\boldsymbol{x}$ is through the quadratic form

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \tag{12}$$

which appears in the exponent. The quantity Δ is called the *Mahalanobis distance* from $\boldsymbol{\mu}$ to $\boldsymbol{x}$ and reduces to the Euclidean distance when $\boldsymbol{\Sigma}$ is the identity matrix.

First of all, the matrix $\boldsymbol{\Sigma}$ can be taken to be symmetric, without loss of generality. Now consider the eigenvector equation for the covariance matrix

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{13}$$

where $i = 1, \ldots, D$. Because $\boldsymbol{\Sigma}$ is real, symmetric matrix its eigenvalues will be real, and its eigenvectors can be chosen to form an orthonormal set

$$\mathbf{u}_i^{\mathrm{T}} \mathbf{u}_j = I_{ij} \tag{14}$$

where I_{ij} is the i, j element of the identify matrix and satisfies

$$I_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

$\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ can be expressed as

$$\begin{aligned} \boldsymbol{\Sigma} &= \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}} \\ \boldsymbol{\Sigma}^{-1} &= \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}} \end{aligned} \tag{16}$$

The quadratic form becomes

$$\begin{aligned} \Delta^2 &= \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \\ y_i &= \mathbf{u}_i^{\mathrm{T}} (\mathbf{x}-\boldsymbol{\mu}) \end{aligned} \tag{17}$$

We can interpret y_i as a new coordinate system defined by the orthonormal vectors $\boldsymbol{\mu}_i$ that are shifted and rotated with respect to the original x_i coordinates. Forming the vector $\boldsymbol{y} = (y_1, \ldots, y_D)^T$, we have

$$\mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu}) \tag{18}$$

where $\mathbf{U}$ is a matrix whose rows are given by $\mathbf{u}_i^{\mathrm{T}}$. It's an orthogonal matrix, i.e., it satisfies $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$, and here also $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

The quadratic form, and hence the Gaussian density, will be constant on surfaces for which Equation 17 is constant. As illustrated in Figure 1, a Gaussian in a two-dimensional space $\boldsymbol{x} = (x_1, x_2)$ is exp(-1/2) of its value at $\boldsymbol{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the

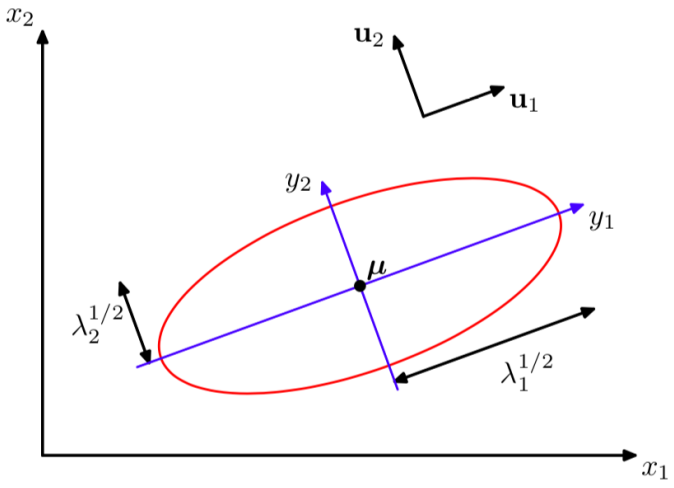

Figure 1: The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space.

eigenvectors $\boldsymbol{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i.

Now consider the form of the Gaussian distribution in the new coordinate system defined by the y_i, a Jacobian matrix $\mathbf{J}$ with elements given by

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji} \tag{19}$$

where U_{ji} are the elements of the matrix $\mathbf{U}^{\mathrm{T}}$.

$$|\mathbf{J}|^2 = \left|\mathbf{U}^{\mathrm{T}}\right|^2 = \left|\mathbf{U}^{\mathrm{T}}\right| |\mathbf{U}| = \left|\mathbf{U}^{\mathrm{T}}\mathbf{U}\right| = |\mathbf{I}| = 1 \tag{20}$$

and hence $|\mathbf{U}| = 1$. The determinant $|\boldsymbol{\Sigma}|$ of the covariance matrix can be written as the product of its eigenvalues, and hence

$$|\boldsymbol{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2} \tag{21}$$

Thus in the y_j coordinate system, the Gaussian distribution takes the form

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\} \tag{22}$$

which is the product of D independent univariate Gaussian distributions. The eigenvectors therefore define a new set of shifted and rotated coordinates with respect to which **the joint probability distribution factorizes into a product of independent distributions**. The integral of the distribution in the $\boldsymbol{y}$ coordinate system is then

$$\int p(\mathbf{y})\mathrm{d}\mathbf{y} = \prod_{j=1}^{D} \int_{-\infty}^{\infty} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\} \mathrm{d}y_j = 1 \tag{23}$$

This confirms that the multivariate Gaussian is indeed normalized.

$$\begin{aligned} \mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu} \\ \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] &= \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} \\ \mathrm{cov}[\mathbf{x}] &= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}] = \boldsymbol{\Sigma} \end{aligned} \tag{24}$$

Consider the number of free parameters in the distribution. A general symmetric covariance matrix $\boldsymbol{\Sigma}$ will have $D(D+1)/2$ independent parameters, and there are another D independent parameters in $\boldsymbol{\mu}$, giving $D(D+3)/2$ parameters in total. For large D, the computation is expensive. One way is to use restricted forms of the covariance matrix. *diagonal* covariance matrices, so that $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$, we then have a total of $2D$ independent parameters in the density model. We could further restrict the covariance matrix to be proportional to the identity matrix, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, known as an *isotropic covariance*, giving $D+1$ independent parameters in the model.

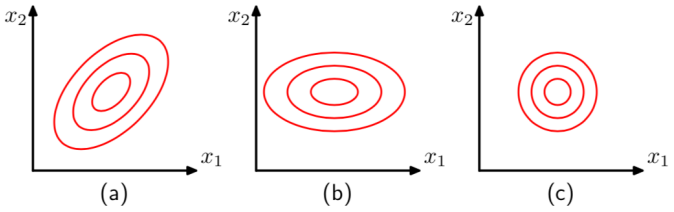

Figure 2: Contours of constant probability density for a Gaussian distribution in which the covariance matrix is (a) of general form, (b) diagonal, agligned with the coordinate axes, and (c), concentric circles.

The three possibilities of general, diagonal, and isotropic covariance matrices are illustrated in Figure 2. Such approaches limit the number of degrees of freedom and make inversion of the covariance matrix a much faster operation, but they also greatly restrict the form of the probability density and limit its ability to capture interesting correlations in the data. A further limitation of the Gaussian distribution is that it is intrinsically unimodal (i.e., has a single maximum) and so is unable to provide a good approximation to multimodal distributions.

3.1. Conditional Gaussian distributions

An important property of the multivariate Gaussian distribution is that if two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is Gaussian, the marginal distribution of either set is also Gaussian.

Suppose $\boldsymbol{x}$ is a D-dimensional vector with Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that we partition $\boldsymbol{x}$ into two disjoint subsets $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, $\boldsymbol{x}_a$ includes the first M components of $\boldsymbol{x}$, and $\boldsymbol{x}_b$ comprises the remaining $D - M$ components.

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} \tag{25}$$

The mean vector μ

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{26}$$

and the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{27}$$

Note that the symmetry $\boldsymbol{\Sigma}^{\mathrm{T}} = \boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{aa}$ and $\boldsymbol{\Sigma}_{bb}$ are symmetric, $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^{\mathrm{T}}$.

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \end{aligned} \tag{28}$$

Note that the mean of the conditional distribution $p(x_a|x_b)$ is a linear function of $\boldsymbol{x}_b$ and that the covariance is independent of $\boldsymbol{x}_a$. This represents an example of a *linear-Gaussian* model.

3.2. Marginal Gaussian distributions

The marginal distribution given of

$$p\left(\mathbf{x}_a\right) = \int p\left(\mathbf{x}_a, \mathbf{x}_b\right) \mathrm{d}\mathbf{x}_b \tag{29}$$

which is also Gaussian. Efficient evaluation will be to identify the mean and covariance of the marginal distribution $p(\mathbf{x}_a)$

$$\begin{aligned} \mathbb{E}\left[\mathbf{x}_a\right] &= \boldsymbol{\mu}_a \\ \mathrm{cov}\left[\mathbf{x}_a\right] &= \boldsymbol{\Sigma}_{aa} \end{aligned} \tag{30}$$

$$p\left(\mathbf{x}_a\right) = \mathcal{N}\left(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}\right) \tag{31}$$

The conditional and marginal distributions associated with a multivariate Gaussian involving two variables are shown in Figure 3.

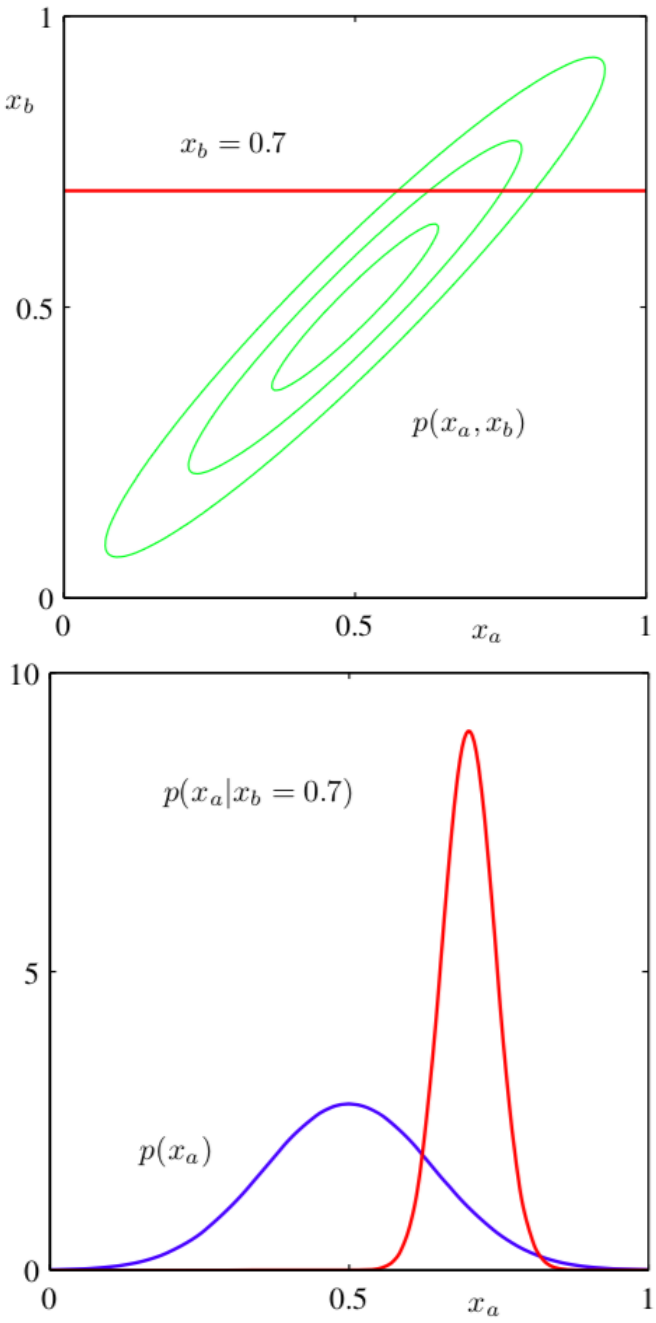

Figure 3: The plot on the top shows the contours of a Gaussian distribution $p(x_a, x_b)$ over two variables, and the plot on the bottom shows the marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve).

3.3. Maximum likelihood for the Gaussian

Given a data set $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T$ in which the observations $\{\boldsymbol{x}_n\}$ are assumed to be drawn independently

from a multivariate Gaussian distribution. The log likelihood function is given by $\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) =$

$$-\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \tag{32}$$

The derivative of the log likelihood with respect to μ is given by

$$\frac{\partial}{\partial \boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \tag{33}$$

and setting this derivative to zero, the maximum likelihood estimate of the mean given by

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{34}$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathrm{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \tag{35}$$

which involves $\boldsymbol{\Sigma}_{\mathrm{ML}}$ because this is the result of a joint maximization with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] &= \boldsymbol{\mu} \\ \mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{N-1}{N}\boldsymbol{\Sigma}\end{aligned} \tag{36}$$

The expectation of the maximum likelihood estimate for the mean is equal to the true mean. However, the maximum likelihood estimate for the covariance has an expectation that is less than the true value, and hence it is biased.

3.4. Bayesian inference for the Gaussian

The maximum likelihood framework gave point estimates for the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Given a set of N observations $\boldsymbol{X}$, we suppose that the variance σ^2 is known, and consider the task of inferring the mean μ.

$$p(\mathbf{X}|\mu) = \prod_{n=1}^{N}p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \tag{37}$$

Note that $p(\boldsymbol{X}|\mu)$ is not probability distribution over μ and is not normalized.

If we choose a prior $p(\mu)$ given by a Gaussian, it will be a conjugate distribution for this likelihood function.

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) \tag{38}$$

and the posterior distribution is given by

$$p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu)p(\mu) \tag{39}$$

$$p(\mu|\mathbf{X}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

where

$$\begin{aligned}\mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}} \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\end{aligned} \tag{40}$$

Now let us suppose that the mean is known and we wish to infer the variance. It turns out to be most convenient to work with the precision $\lambda \equiv 1/\sigma^2$. The likelihood function for λ takes the form $p(\mathbf{X}|\lambda) =$

$$\prod_{n=1}^{N}\mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2}\exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \tag{41}$$

This corresponds to the *gamma* distribution which is defined by

$$\mathrm{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda) \tag{42}$$

The gamma distribution has a finite integral if $a > 0$, and the distribution itself is finite if $a \geq 1$. The mean and variance of the gamma distribution are given by

$$\begin{aligned}\mathbb{E}[\lambda] &= \frac{a}{b} \\ \mathrm{var}[\lambda] &= \frac{a}{b^2}\end{aligned} \tag{43}$$

Consider a prior distribution $\mathrm{Gam}(\lambda|a_0, b_0)$. If we multiply by the likelihood function (41), then we obtain a posterior distribution

$$p(\lambda|\mathbf{X}) \propto \lambda^{a_0-1}\lambda^{N/2}\exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \tag{44}$$

which we recognize as a gamma distribution of the form $\mathrm{Gam}(\lambda|a_N, b_N)$ where

$$\begin{aligned}a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 = b_0 + \frac{N}{2}\sigma_{\mathrm{ML}}^2\end{aligned} \tag{45}$$

where σ_{ML}^2 is the maximum likelihood estimator of the variance.

3.5. Student's t-distribution

If we have a univariate Gaussian $\mathcal{N}(x|\mu, \tau^{-1})$ together with a Gamma prior $\mathrm{Gam}(\tau|a, b)$ and we integrate out the precision, we obtain the marginal distribution of x in the form

$$\begin{aligned} p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}\left(x|\mu, \tau^{-1}\right) \mathrm{Gam}(\tau|a, b)\mathrm{d}\tau \\ &= \int_0^\infty \frac{b^a e^{(-b\tau)}\tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\} \mathrm{d}\tau \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2} \Gamma(a + 1/2) \end{aligned} \tag{46}$$

By convention we define new parameters given by $\nu = 2a$ and $\lambda = a/b$, in terms of which the distribution $p(x|\mu, \lambda, \nu)$ takes the form

$$\mathrm{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2} \tag{47}$$

which is known as *Student's t-distribution*. The parameter λ is sometimes called the *precision* of the t-distribution, the parameters ν is called the *degrees of freedom*.

From (46), we see that Student's t-distribution is obtained by adding up an infinite number of Gaussian distributions having the same mean but different precisions. This can be interpreted as an infinite mixture of Gaussians. This gives the t-distribution an important property called *robustness*, which means that it's much less sensitive than the Gaussian to the presence of a few data points which are *outliers*, as was seen in Figure 4.

Note that the maximum likelihood solution for the t-distribution can be found using the expectation-maximization (EM) algorithm. Outliers can arise in practical applications either because the process that generates the data corresponds to a distribution having a heavy tail or simply through mislabelled data. The least squares approach to regression does not exhibit robustness, because it corresponds to maximum likelihood under a (conditional) Gaussian distribution. By basing a regression model on a heavy-tailed distribution such as t-distribution, we obtain a more robust model.

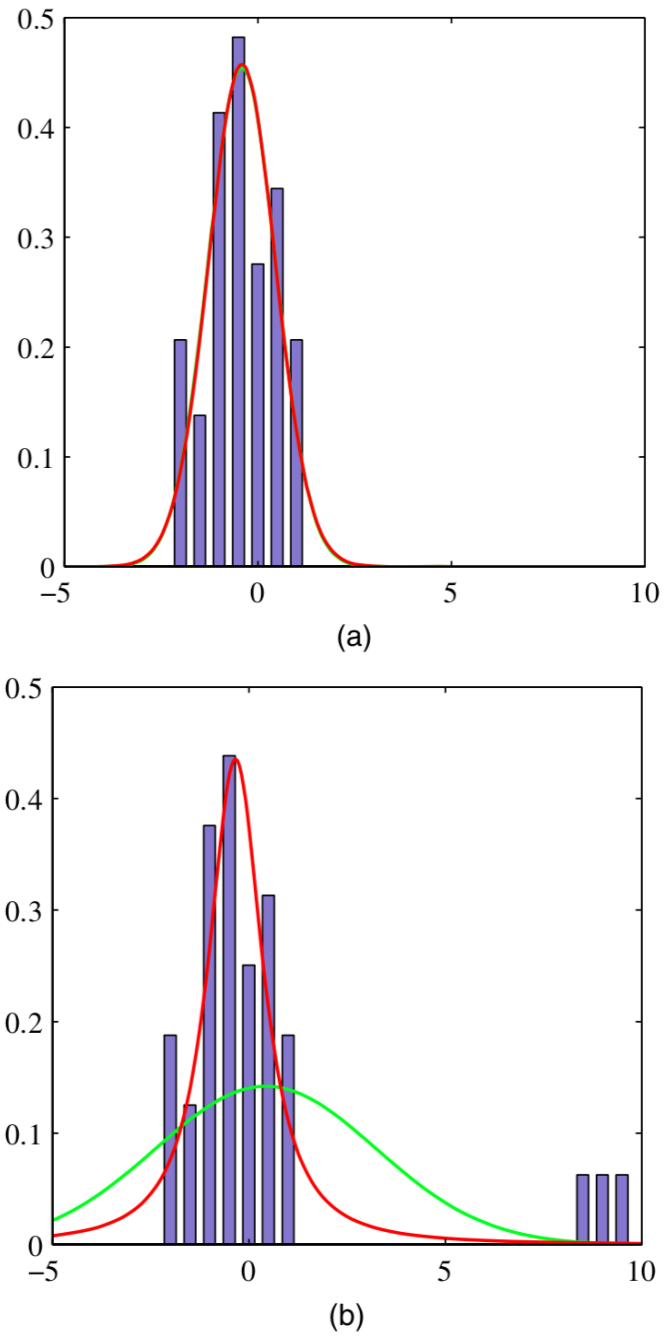

Figure 4: Illustration of robustness of Student's t-distribution compared to a Gaussian. (a) Histogram distribution of 30 data points drawn from a Gaussian distribution, together with the maximum likelihood fit obtained from a t-distribution (red curve) and a Gaussian (green curve, largely hidden by the red curve). Because the t-distribution contains the Gaussian as a special case it gives almost the same solution as the Gaussian. (b) The same data set but with three additional outlying data points showing how the Gaussian (green curve) is strongly distorted by the outliers, whereas the t-distribution (red curve) is relatively unaffected.

3.6. Mixtures of Gaussians

In Figure 5 we see that a linear combination of Gaussians can give rise to very complex densities. By using

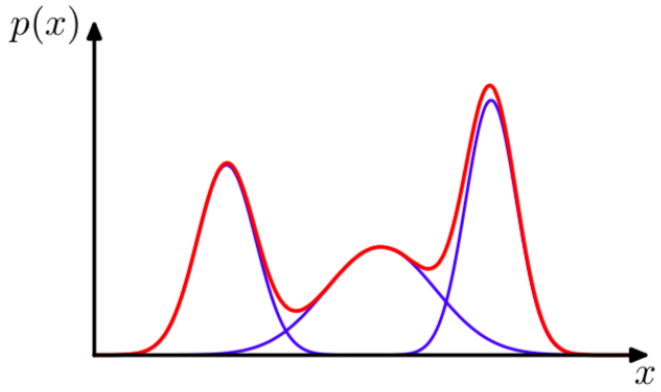

Figure 5: Example of a Gaussian mixture distribution in one dimension showing three Gaussians (each scaled by a coefficient) in blue and their sum in red.

a sufficient number of Gaussians, ans by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy.

We therefore consider a superposition of K Gaussian densities of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{48}$$

which is called a *mixture of Gaussians*. Each Gaussian density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is called a *component* of the mixture and has its own mean $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. The parameters π_k are called *mixing coefficients*.

$$\begin{aligned} &\sum_{k=1}^{K} \pi_k = 1, \\ &0 \leqslant \pi_k \leqslant 1 \end{aligned} \tag{49}$$

From the sum and product rules, the marginal density is given by

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k) \tag{50}$$

we can view $\pi_k = p(k)$ as the prior probability of picking the k^{th} component, and the density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\mathbf{x}|k)$ as the probability of $\boldsymbol{x}$ conditioned on k.

4. The Exponential Family

The *exponential family* of distributions over $\boldsymbol{x}$, given parameters $\boldsymbol{\eta}$, is defined to be set of distributions of the form

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \tag{51}$$

where $\boldsymbol{x}$ may be scalar or vector, and may be discrete or continuous. Here $\boldsymbol{\eta}$ are called *natural parameters* of the distribution, and $\mathbf{u}(\mathbf{x})$ is some function of $\mathbf{x}$. The function $g(\eta)$ can be interpreted as the coefficient that ensures that the distribution is normalized and therefore satisfies

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathrm{d}\mathbf{x} = 1 \tag{52}$$

where the integration is replaced by summation if $\mathbf{x}$ is a discrete variable.

Consider first the Bernoulli distribution

$$p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} \tag{53}$$

which $\mu = \sigma(\eta)$, σ is the logistic sigmoid function.

$$p(x|\eta) = \sigma(-\eta) \exp(\eta x) \tag{54}$$

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= \sigma(-\eta) \end{aligned} \tag{55}$$

Next consider the multinomial distribution

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} \tag{56}$$

where $\mathbf{x} = (x_1, \ldots, x_N)^T$, $\eta_k = \ln \mu_k$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^T$.

$$p(\mathbf{x}|\boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{x}\right) \tag{57}$$

Note that the parameters η_k are not independent because the parameters μ_k are subject to the constraint

$$\sum_{k=1}^{M} \mu_k = 1 \tag{58}$$

$$\begin{aligned} &\exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} \\ &= \exp\left\{\sum_{k=1}^{M-1} x_k \ln \mu_k + \left(1 - \sum_{k=1}^{M-1} x_k\right) \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\} \\ &= \exp\left\{\sum_{k=1}^{M-1} x_k \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) + \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\} \end{aligned} \tag{59}$$

$$\begin{aligned} \ln\left(\frac{\mu_k}{1 - \sum_j \mu_j}\right) &= \eta_k \\ \mu_k &= \frac{\exp\left(\eta_k\right)}{1 + \sum_j \exp\left(\eta_j\right)} \end{aligned} \tag{60}$$

which is called the *softmax* function.

$$p(\mathbf{x}|\boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp\left(\eta_k\right)\right)^{-1} \exp\left(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{x}\right) \tag{61}$$

This is the standard form of the exponential family, with parameter vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{M-1})^T$ in which

$$\begin{aligned} \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= \left(1 + \sum_{k=1}^{M-1} \exp\left(\eta_k\right)\right)^{-1} \end{aligned} \tag{62}$$

5. Nonparametric Methods

The probability distributions having specific functional forms are governed by a small number of parameters whose values are to be determined from a data set. This is called the *parametric* approach to density modelling. An important limitation of this approach is that the chosen density might be a poor model of the distribution that generates the data.

Here, we consider some *nonparametric* approaches to density estimation that make few assumptions about the form of the distribution. Let us start with a discussion of histogram methods for density estimation.

Given a single continuous variable x, standard histograms simply partition x into distinct bins of width Δ and then count the number n_i of observations of x falling in bin i.

$$p_i = \frac{n_i}{N\Delta} \tag{63}$$

for which it is easily seen that $\int p(x)\mathrm{d}x = 1$. This gives a model for the density $p(x)$ that is constant over the width of each bin.

In Figure 6, we see that when Δ is very small (top figure), the resulting density model is very spiky, with a lot of structure that is not present in the underlying distribution that generated the data set. If Δ is too large (bottom figure) then the result is a model that is too smooth and that consequently fails to capture the bimodal property of the green curve. The best results are obtained for some intermediate value of Δ (middle figure).

Note that the histogram method has the property that, once the histogram has been computed, the data set itself

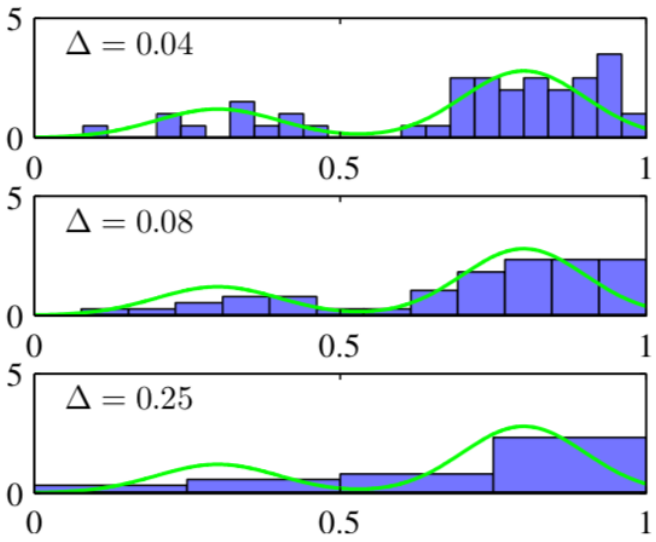

Figure 6: An illustration of the histogram approach to density estimation, in which a data set of 50 data points is generated from the distribution shown by the green curve, which is formed from a mixture of two Gaussians.

can be discarded, which can be advantageous if the data set is large. Also the histogram approach is easily applied if the data points are arriving sequentially.

References

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

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