CHAPTER 2: PROBABILITY DISTRIBUTIONS

Binary Variables (1)

Coin flipping: heads=1, tails=0

\[ p(x = 1|\mu) = \mu \]

Bernoulli Distribution

\[
\begin{align*}
\text{Bern}(x|\mu) &= \mu^x (1 - \mu)^{1-x} \\
\mathbb{E}[x] &= \mu \\
\text{var}[x] &= \mu(1 - \mu)
\end{align*}
\]

Binary Variables (2)

\( N \) coin flips:

\[ p(x \text{ heads}|N, \mu) \]

Binomial Distribution

\[
\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}
\]

\[
\mathbb{E}[m] = \sum_{m=0}^{N} m \text{Bin}(m|N, \mu) = N\mu
\]

\[
\text{var}[m] = \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N\mu(1 - \mu)
\]

Binomial Distribution

![Binomial Distribution Graph](image)
The Multinomial Distribution

Multinomial distribution is a generalization of the binomial distribution. Different from the binomial distribution, where the RV assumes two outcomes, the RV for multinominal distribution can assume \( k \) (>2) possible outcomes.

Let \( N \) be the total number of independent trials, \( m_i, i=1,2,..,k \), be the number of times outcome \( i \) appears. Then, performing \( N \) independent trials, the probability that outcome 1 appears \( m_1 \), outcome 2, appears \( m_2 \), …, outcome \( k \) appears \( m_k \) times is

\[
\operatorname{Mult}(m_1, m_2, \ldots, m_k; \mu, N) = \binom{N}{m_1, m_2, \ldots, m_k} \prod_{i=1}^{k} \mu_i^{m_i}
\]

\[
\mathbb{E}[m_i] = N \mu_i
\]

\[
\text{var}(m_i) = N \mu_i(1 - \mu_i)
\]

\[
\text{cov}(m_i, m_k) = -N \mu_i \mu_k
\]

The Gaussian Distribution

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

Moments of the Multivariate Gaussian (1)

\[
\mathbb{E}[x] = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \int \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\} x \, dx
\]

\[
= \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \int \exp \left\{ -\frac{1}{2} z^\top \Sigma^{-1} z \right\} (z + \mu) \, dz
\]

thanks to anti-symmetry of \( \Sigma \)

\[
\mathbb{E}[x] = \mu
\]

Moments of the Multivariate Gaussian (2)

\[
\mathbb{E}[xx^\top] = \mu \mu^\top + \Sigma
\]

\[
\text{cov}[x] = \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^\top] = \Sigma
\]
Central Limit Theorem

The distribution of the sum of \(N\) i.i.d. random variables becomes increasingly Gaussian as \(N\) grows.

Example: \(N\) uniform \([0, 1]\) random variables.

Beta Distribution

Beta is a continuous distribution defined on the interval of 0 and 1, i.e., \(\mu \in [0, 1]\) parameterized by two positive parameters \(a\) and \(b\).

\[
\begin{align*}
\text{Beta}(\mu | a, b) &= \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \\
\mathbb{E}[\mu] &= \frac{a}{a + b} \\
\text{var}[\mu] &= \frac{ab}{(a + b)^2(a + b + 1)}
\end{align*}
\]

where \(\Gamma(\cdot)\) is gamma function. beta is conjugate to the binomial and Bernoulli distributions.

The Dirichlet Distribution

The Dirichlet distribution is a continuous multivariate probability distribution parametrized by a vector of positive reals \(\alpha\). It is the multivariate generalization of the beta distribution.

\[
\text{Dir}(\mu | \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k-1}
\]

\[\alpha_0 = \sum_{k=1}^{K} \alpha_k\]

Conjugate prior for the multinomial distribution.
Mixtures of Gaussians (1)

Old Faithful data set

Mixtures of Gaussians (2)

Combine simple models into a complex model:

\[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k) \]

where \( \pi_k \geq 0 \) and \( \sum_{k=1}^{K} \pi_k = 1 \)

Mixtures of Gaussians (3)

The Exponential Family (1)

\[ p(x|\eta) = h(x) g(\eta) \exp \{ \eta^T u(x) \} \]

where \( \eta \) is the natural parameter and

\[ g(\eta) \int h(x) \exp \{ \eta^T u(x) \} \, dx = 1 \]

so \( g(\eta) \) can be interpreted as a normalization coefficient.
The Exponential Family (2.1)

The Bernoulli Distribution

\[ p(x|\mu) = \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \]
\[ = \exp\{x\ln\mu + (1-x)\ln(1-\mu)\} \]
\[ = (1-\mu)\exp\{\ln\left(\frac{\mu}{1-\mu}\right)x\} \]

Comparing with the general form we see that

\[ \eta = \ln\left(\frac{\mu}{1-\mu}\right) \quad \text{and so} \quad \mu = \sigma(\eta) = \frac{1}{1 + \exp(-\eta)} \]

Logistic sigmoid

The Exponential Family (2.2)

The Bernoulli distribution can hence be written as

\[ p(x|\eta) = \sigma(-\eta)\exp(\eta x) \]

where

\[ u(x) = x \]
\[ h(x) = 1 \]
\[ g(\eta) = 1 - \sigma(\eta) = \sigma(-\eta). \]

The Exponential Family (3.1)

The Multinomial Distribution

\[ p(x|\mu) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{ \sum_{k=1}^{M} x_k \ln \mu_k \right\} = h(x)g(\eta)\exp(\eta^Tu(x)) \]

where, \( x = (x_1, \ldots, x_M)^T \), \( \eta = (\eta_1, \ldots, \eta_M)^T \) and

\[ \eta_k = \ln \mu_k \]
\[ u(x) = x \]
\[ h(x) = 1 \]
\[ g(\eta) = 1. \]

NOTE: The \( \mu_k \) parameters are not independent since the corresponding \( \mu_k \) must satisfy \( \sum_{k=1}^{M} \mu_k = 1 \).

The Exponential Family (3.2)

Let \( \mu_M = 1 - \sum_{k=1}^{M-1} \mu_k \). This leads to

\[ \eta_k = \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) \quad \text{and} \quad \mu_k = \frac{\exp(\eta_k)}{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}. \]

Softmax

Here the \( \eta_k \) parameters are independent. Note that

\[ 0 \leq \mu_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{M} \mu_k = 1. \]
The Exponential Family (3.3)

The Multinomial distribution can then be written as

\[ p(\mathbf{x} | \mathbf{\mu}) = h(\mathbf{x}) g(\mathbf{\eta}) \exp \left( \mathbf{\eta}^T \mathbf{u}(\mathbf{x}) \right) \]

where

\[ \eta = (\eta_1, \ldots, \eta_{M-1}, 0)^T \]
\[ u(x) = x \]
\[ h(x) = 1 \]
\[ g(\eta) = \left( 1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right)^{-1} \]

The Exponential Family (4)

The Gaussian Distribution

\[ p(x | \mu, \sigma^2) = \frac{1}{(2\pi \sigma^2)^{1/2}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right) \]
\[ = \frac{1}{(2\pi \sigma^2)^{1/2}} \exp \left( -\frac{1}{2\sigma^2} x^2 + \frac{\mu^2}{\sigma^2} x - \frac{1}{2\sigma^2} \mu^2 \right) \]
\[ = h(x) g(\eta) \exp \left( \eta^T \mathbf{u}(x) \right) \]

where

\[ \eta = \left( \frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right) \]
\[ h(x) = (2\pi)^{-1/2} \]
\[ u(x) = \left( \frac{x^2}{2 \sigma^2} \right) \]
\[ g(\eta) = (-2\eta_2)^{1/2} \exp \left( \frac{\eta_1^2}{4\eta_2} \right) \]

Conjugate priors

For any member of the exponential family, there exists a prior

\[ p(\mathbf{\eta} | \mathbf{\chi}, \nu) = f(\mathbf{\chi}, \nu) g(\mathbf{\eta})^\nu \exp \left\{ \nu \mathbf{\eta}^T \mathbf{\chi} \right\} \]

Combining with the likelihood function, we get posterior

\[ p(\mathbf{\eta} | \mathbf{X}, \mathbf{\chi}, \nu) \propto g(\mathbf{\eta})^{\nu + N} \exp \left\{ \eta^T \left( \sum_{n=1}^{N} \mathbf{u}(x_n) + \nu \mathbf{\chi} \right) \right\} \]

Conjugate priors (cont’d)

- Beta prior is conjugate to the binomial and Bernoulli distributions
- Dirichlet prior is conjugate to the multinomial distribution
- Gaussian prior is conjugate to the Gaussian distribution
Noninformative Priors (1)

With little or no information available a-priori, we might choose uniform prior.

- $\lambda$ discrete, $K$-nomial: $p(\lambda) = 1/K$.
- $\lambda \in [a,b]$ real and bounded: $p(\lambda) = 1/(b-a)$.
- $\lambda$ real and unbounded: improper!

Nonparametric Methods (1)

Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.

Nonparametric approaches make few assumptions about the overall shape of the distribution being modelled.

Nonparametric Methods (2)

Histogram methods partition the data space into distinct bins with widths $\Delta_i$ and count the number of observations, $n_i$, in each bin.

$$p_i = \frac{n_i}{N\Delta_i}$$

- Often, the same width is used for all bins, $\Delta_i = \Delta$.
- $\Delta$ acts as a smoothing parameter.

Nonparametric Methods (3)

Kernel Density Estimation: is a non-parametric way of estimating the probability density function of a random variable.

Let $(x_1, x_2, ..., x_n)$ be an iid sample drawn from some distribution with an unknown density $p(x)$ (Parzen window)

$$k\left(\frac{x-x_i}{h}\right) = \begin{cases} \frac{1}{h} & \text{if } |x-x_i| < h/2 \\ 0, \text{ else} \end{cases}$$

It follows that

$$p(x) = \frac{1}{Nh} \sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right)$$

$k(\cdot)$ is the kernel function and $h$ is bandwidth, serving as a smoothing parameter. The only parameter is $h$. 
Nonparametric Methods (4)

To avoid discontinuities in \( p(x) \), use a smooth kernel, e.g. a Gaussian

\[
p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{D/2}} \exp \left( -\frac{||x - x_n||^2}{2h^2} \right)
\]

Any kernel such that

\[
\int k(x) \, dx = 1
\]

will work.

Nonparametric Methods (5)

Nonparametric models (not histograms) requires storing and computing with the entire data set.

Parametric models, once fitted, are much more efficient in terms of storage and computation.

\[ K \text{-Nearest-Neighbours for Classification} \]

The k-nearest neighbors algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space.

The best choice of \( k \) depends upon the data; larger values of \( k \) reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good \( k \) can be selected by cross-validation.
**K-Nearest-Neighbours for Classification (3)**

- K acts as a smoother
- For $N \to \infty$, the error rate of the 1-nearest-neighbour classifier is never more than twice the optimal error (obtained from the true conditional class distributions).

---

**Parametric Estimation**

**Basic building blocks:** $p(x|\theta)$

Need to determine $\theta$ given $\{x_1, \ldots, x_N\}$

**Maximum Likelihood (ML)**

$$\theta^* = \arg \max_\theta p(x_1, x_2, \ldots, x_n | \theta)$$

**Maximum Posterior Probability (MAP)**

$$\theta^* = \arg \max_\theta p(\theta | x_1, x_2, \ldots, x_n)$$

---

**ML Parameter Estimation**

Since samples $x_1, x_2, \ldots, x_n$ are i.i.d., we have

$$L(\theta) = p(x_1, x_2, \ldots, x_n | \theta)$$

The log likelihood can be obtained as

$$\log L(\theta) = \sum p(x_n | \theta)$$

$\theta$ can be obtained by taking the derivative of the Log likelihood with respect to $\theta$ and setting it to zero

$$\frac{\partial \log L(\theta)}{\partial \theta} = \sum \frac{\partial p(x_n | \theta)}{\partial \theta} = 0$$

---

**Parameter Estimation (1)**

**ML for Bernoulli**

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, $m$ heads (1), $N - m$ tails (0)

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n | \mu) = \prod_{x_n=0}^{K} \mu^{x_n} (1-\mu)^{1-x_n}$$

$$\ln p(D|\mu) = \sum_{n=1}^{N} \ln p(x_n | \mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n) \ln(1-\mu)\}$$

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N}$$
Maximum Likelihood for the Gaussian

Given i.i.d. data $\mathbf{x} = (x_1, \ldots, x_N)^T$, the log likelihood function is given by

$$\ln p(\mathbf{x} | \mu, \Sigma) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)\Sigma^{-1}(x_n - \mu)$$

Sufficient statistics

$$\sum_{n=1}^{N} x_n \quad \sum_{n=1}^{N} x_n x_n^T$$

Set the derivative of the log likelihood function to zero, and solve to obtain

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

Similarly

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T.$$
Bayesian Inference for the Gaussian (2)

Combined with a Gaussian prior over $\mu$,

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

this gives the posterior

$$p(\mu | x) \propto p(x | \mu)p(\mu).$$

Completing the square over $\mu$, we see that

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

Bayesian Inference for the Gaussian (3)

... where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML},$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}. $$

Note:

<table>
<thead>
<tr>
<th>$\mu_N$</th>
<th>$\mu_0$</th>
<th>$\mu_{ML}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_N^2$</td>
<td>$\sigma_0^2$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

Bayesian Inference for the Gaussian (4)

Example: $p(\mu | x) = \mathcal{N}(\mu | \mu_N, \sigma_N^2)$ for $N = 0, 1, 2$ and 10.