10707
Deep Learning
Russ Salakhutdinov
Machine Learning Department
rsalakhu@cs.cmu.edu
http://www.cs.cmu.edu/~rsalakhu/10707/

Lecture 3
Bernoulli Distribution

• Consider a single binary random variable $x \in \{0, 1\}$. For example, $x$ can describe the outcome of flipping a coin:

  Coin flipping: heads = 1, tails = 0.

• The probability of $x=1$ will be denoted by the parameter $\mu$, so that:

  $$p(x = 1 | \mu) = \mu \quad 0 \leq \mu \leq 1.$$  

• The probability distribution, known as Bernoulli distribution, can be written as:

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

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

  $$\text{var}[x] = \mu(1 - \mu)$$
Parameter Estimation

• Suppose we observed a dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$

• We can construct the likelihood function, which is a function of $\mu$.

\[ p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n} (1 - \mu)^{1-x_n} \]

• Equivalently, we can maximize the log of the likelihood function:

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

• Note that the likelihood function depends on the N observations $x_n$ only through the sum $\sum_n x_n$.
Parameter Estimation

• Suppose we observed a dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$

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

• Setting the derivative of the log-likelihood function w.r.t $\mu$ to zero, we obtain:

\[
\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N}
\]

where $m$ is the number of heads.
Binomial Distribution

• We can also work out the distribution of the number m of observations of x=1 (e.g. the number of heads).

• The probability of observing m heads given N coin flips and a parameter $\mu$ is given by:

$$p(m \text{ heads}|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

• The mean and variance can be easily derived as:

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m \binom{N}{m} \mu^m (1 - \mu)^{N-m} = N \mu$$

$$\text{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \binom{N}{m} \mu^m (1 - \mu)^{N-m} = N \mu (1 - \mu)$$
Example

• Histogram plot of the Binomial distribution as a function of $m$ for $N=10$ and $\mu = 0.25$. 

$\text{Bin}(m|10, 0.25)$
Beta Distribution

- We can define a distribution over $\mu \in [0, 1]$ (e.g. it can be used a prior over the parameter $\mu$ of the Bernoullli distribution).

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

$$
\mathbb{E}[\mu] = \frac{a}{a + b}
$$

$$
\text{var}[\mu] = \frac{ab}{(a + b)^2(a + b + 1)}
$$

where the gamma function is defined as:

$$
\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} du.
$$

and ensures that the Beta distribution is normalized.
Beta Distribution

\[ a = 0.1 \]
\[ b = 0.1 \]

\[ a = 1 \]
\[ b = 1 \]

\[ a = 2 \]
\[ b = 3 \]

\[ a = 8 \]
\[ b = 4 \]
Multinomial Variables

- Consider a random variable that can take on one of K possible mutually exclusive states (e.g. roll of a dice).

- We will use so-called 1-of-K encoding scheme.

- If a random variable can take on K=6 states, and a particular observation of the variable corresponds to the state $x_3=1$, then $\mathbf{x}$ will be resented as:

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

1-of-K coding scheme:

- If we denote the probability of $x_k=1$ by the parameter $\mu_k$, then the distribution over $\mathbf{x}$ is defined as:

$$p(\mathbf{x} | \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \quad \forall k : \mu_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$
Multinomial Variables

- Multinomial distribution can be viewed as a generalization of Bernoulli distribution to more than two outcomes.

\[ p(x | \mu) = \prod_{k=1}^{K} \mu_k^{x_k} \]

- It is easy to see that the distribution is normalized:

\[ \sum_{x} p(x | \mu) = \sum_{k=1}^{K} \mu_k = 1 \]

and

\[ \mathbb{E}[x | \mu] = \sum_{x} p(x | \mu) x = (\mu_1, \ldots, \mu_K)^T = \mu \]
Maximum Likelihood Estimation

• Suppose we observed a dataset \( \mathcal{D} = \{x_1, \ldots, x_N\} \)

• We can construct the likelihood function, which is a function of \( \mu \).

\[
p(\mathcal{D} | \mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}
\]

• Note that the likelihood function depends on the \( N \) data points only though the following \( K \) quantities:

\[
m_k = \sum_n x_{nk}, \quad k = 1, \ldots, K.
\]

which represents the number of observations of \( x_k=1 \).

• These are called the sufficient statistics for this distribution.
Maximum Likelihood Estimation

\[ p(D|\mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_{k}^{x_{nk}} = \prod_{k=1}^{K} \mu_{k}^{(\sum_{n} x_{nk})} = \prod_{k=1}^{K} \mu_{k}^{m_{k}} \]

- To find a maximum likelihood solution for \( \mu \), we need to maximize the log-likelihood taking into account the constraint that \( \sum_{k} \mu_{k} = 1 \).

- Forming the Lagrangian:

\[
\sum_{k=1}^{K} m_{k} \ln \mu_{k} + \lambda \left( \sum_{k=1}^{K} \mu_{k} - 1 \right)
\]

\[
\mu_{k} = -m_{k}/\lambda \quad \mu_{k}^{\text{ML}} = \frac{m_{k}}{N} \quad \lambda = -N
\]

which is the fraction of observations for which \( x_{k}=1 \).
Multinomial Distribution

• We can construct the joint distribution of the quantities \{m_1, m_2, \ldots, m_k\} given the parameters \(\mu\) and the total number \(N\) of observations:

\[
\text{Mult}(m_1, m_2, \ldots, m_K|\mu, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}
\]

\[
\mathbb{E}[m_k] = N \mu_k
\]

\[
\text{var}[m_k] = N \mu_k (1 - \mu_k)
\]

\[
\text{cov}[m_j m_k] = -N \mu_j \mu_k
\]

• The normalization coefficient is the number of ways of partitioning \(N\) objects into \(K\) groups of size \(m_1, m_2, \ldots, m_k\).

• Note that

\[
\sum_k m_k = N.
\]
Dirichlet Distribution

• Consider a distribution over $\mu_k$, subject to constraints:

\[ \forall k : \mu_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1 \]

• The Dirichlet distribution is defined as:

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

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

where $\alpha_1, \ldots, \alpha_k$ are the parameters of the distribution, and $\Gamma(x)$ is the gamma function.

• The Dirichlet distribution is confined to a simplex as a consequence of the constraints.
Dirichlet Distribution

- Plots of the Dirichlet distribution over three variables.

\[ \alpha_k = 10^{-1} \quad \alpha_k = 10^0 \quad \alpha_k = 10^1 \]
Gaussian Univariate Distribution

• In the case of a single variable $x$, the Gaussian distribution takes form:

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

which is governed by two parameters:

- $\mu$ (mean)
- $\sigma^2$ (variance)

• The Gaussian distribution satisfies:

$$
\mathcal{N}(x|\mu, \sigma^2) > 0
$$

$$
\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, dx = 1
$$
Multivariate Gaussian Distribution

- For a D-dimensional vector \( \mathbf{x} \), the Gaussian distribution takes form:

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

which is governed by two parameters:

- \( \mu \) is a D-dimensional mean vector.
- \( \Sigma \) is a D by D covariance matrix.

and \( |\Sigma| \) denotes the determinant of \( \Sigma \).

- Note that the covariance matrix is a symmetric positive definite matrix.
Central Limit Theorem

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

• Consider N variables, each of which has a uniform distribution over the interval [0,1].

• Let us look at the distribution over the mean:

\[
\frac{x_1 + x_2 + \ldots + x_N}{N}.
\]

• As N increases, the distribution tends towards a Gaussian distribution.
Moments of the Gaussian Distribution

• The expectation of \( \mathbf{x} \) under the Gaussian distribution:

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

= \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \int \exp \left\{ -\frac{1}{2} \mathbf{z}^T \Sigma^{-1} \mathbf{z} \right\} (\mathbf{z} + \mu) \, d\mathbf{z}

The term in \( \mathbf{z} \) in the factor \( (\mathbf{z}+\mu) \) will vanish by symmetry.

\[
\mathbb{E}[\mathbf{x}] = \mu
\]
Moments of the Gaussian Distribution

• The second order moments of the Gaussian distribution:

\[ \mathbb{E}[\mathbf{x}\mathbf{x}^T] = \mu\mu^T + \Sigma \]

• The covariance is given by:

\[ \text{cov}[\mathbf{x}] = \mathbb{E} \left[ (\mathbf{x} - \mathbb{E}[\mathbf{x}]) (\mathbf{x} - \mathbb{E}[\mathbf{x}])^T \right] = \Sigma \]

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

• Because the parameter matrix \( \Sigma \) governs the covariance of \( \mathbf{x} \) under the Gaussian distribution, it is called the covariance matrix.
Moments of the Gaussian Distribution

- Contours of constant probability density:

\( x_2 \)

(a) Covariance matrix is of general form.

(b) Diagonal, axis-aligned covariance matrix.

(c) Spherical (proportional to identity) covariance matrix.
Partitioned Gaussian Distribution

• Consider a D-dimensional Gaussian distribution:
\[ p(x) = \mathcal{N}(x|\mu, \Sigma) \]

• Let us partition \( x \) into two disjoint subsets \( x_a \) and \( x_b \):

\[
x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}
\]

• In many situations, it will be more convenient to work with the precision matrix (inverse of the covariance matrix):

\[
\Lambda \equiv \Sigma^{-1} \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}
\]

• Note that \( \Lambda_{aa} \) is not given by the inverse of \( \Sigma_{aa} \).
Conditional Distribution

- It turns out that the conditional distribution is also a Gaussian distribution:

\[ p(x_a | x_b) = \mathcal{N}(x_a | \mu_{a|b}, \Sigma_{a|b}) \]

Covariance does not depend on \( x_b \).

\[
\Sigma_{a|b} = \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}
\]

\[
\mu_{a|b} = \Sigma_{a|b} \left\{ \Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b) \right\}
\]

\[
= \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)
\]

\[
= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)
\]

Linear function of \( x_b \).
Marginal Distribution

• It turns out that the marginal distribution is also a Gaussian distribution:

\[
p(x_a) = \int p(x_a, x_b) \, dx_b
\]

\[
= \mathcal{N}(x_a | \mu_a, \Sigma_{aa})
\]

• For a marginal distribution, the mean and covariance are most simply expressed in terms of partitioned covariance matrix.

\[
x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}
\]

\[
\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}
\]

\[
\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}
\]
Conditional and Marginal Distributions

\[ p(x_a, x_b) \]

\[ p(x_a | x_b = 0.7) \]

\[ x_b = 0.7 \]
Maximum Likelihood Estimation

• Suppose we observed i.i.d data $\mathbf{X} = \{x_1, \ldots, x_N\}$.

• We can construct the log-likelihood function, which is a function of $\mu$ and $\Sigma$:

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

• Note that the likelihood function depends on the N data points only though the following sums:

**Sufficient Statistics**

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

$$\sum_{n=1}^{N} x_n x_n^T$$
Maximum Likelihood Estimation

• To find a maximum likelihood estimate of the mean, we set the derivative of the log-likelihood function to zero:

\[
\frac{\partial}{\partial \mu} \ln p(X|\mu, \Sigma) = \sum_{n=1}^{N} \Sigma^{-1}(x_n - \mu) = 0
\]

and solve to obtain:

\[
\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n.
\]

• Similarly, we can find the ML estimate of \( \Sigma \):

\[
\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T.
\]
### Maximum Likelihood Estimation

- Evaluating the expectation of the ML estimates under the true distribution, we obtain:

  
  \[ 
  \mathbb{E}[\mu_{ML}] = \mu \\
  \mathbb{E}[\Sigma_{ML}] = \frac{N - 1}{N} \Sigma. 
  \]

  - Note that the maximum likelihood estimate of \( \Sigma \) is biased.

- We can correct the bias by defining a different estimator:

  \[ 
  \tilde{\Sigma} = \frac{1}{N - 1} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T. 
  \]
Student’s t-Distribution

• Consider Student’s t-Distribution

\[
p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) \, d\tau
\]

\[
= \int_0^\infty \mathcal{N}(x|\mu, (\eta\lambda)^{-1}) \text{Gam}(\eta|\nu/2, \nu/2) \, d\eta
\]

\[
= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi \nu}\right)^{1/2} \left[1 + \frac{\lambda(x - \mu)^2}{\nu}\right]^{-\nu/2 - 1/2}
\]

\[
= \text{St}(x|\mu, \lambda, \nu)
\]

where

\[
\lambda = a/b \quad \eta = \tau b/a \quad \nu = 2a.
\]

Sometimes called the precision parameter.

Degrees of freedom

Infinite mixture of Gaussians
Student’s t-Distribution

- Setting $\nu = 1$ recovers Cauchy distribution
- The limit $\nu \to \infty$ corresponds to a Gaussian distribution.

<table>
<thead>
<tr>
<th>$\nu = 1$</th>
<th>$\nu \to \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>St$(x</td>
<td>\mu, \lambda, \nu)$</td>
</tr>
</tbody>
</table>

\[ \nu \to \infty \]
\[ \nu = 1.0 \]
\[ \nu = 0.1 \]
Student’s t-Distribution

• Robustness to outliers: Gaussian vs. t-Distribution.
Student’s t-Distribution

• The multivariate extension of the t-Distribution:

\[
\text{St}(x|\mu, \Lambda, \nu) = \int_0^\infty \mathcal{N}(x|\mu, (\eta \Lambda)^{-1}) \text{Gam}(\eta|\nu/2, \nu/2) \, d\eta
\]

\[
= \frac{\Gamma(D/2 + \nu/2)}{\Gamma(\nu/2)} \frac{|\Lambda|^{1/2}}{(\pi \nu)^{D/2}} \left[ 1 + \frac{\Delta^2}{\nu} \right]^{-D/2-\nu/2}
\]

where \(\Delta^2 = (x - \mu)^T \Lambda (x - \mu)\)

• Properties:

\[\mathbb{E}[x] = \mu, \quad \text{if } \nu > 1\]

\[\text{cov}[x] = \frac{\nu}{(\nu - 2)} \Lambda^{-1}, \quad \text{if } \nu > 2\]

\[\text{mode}[x] = \mu\]
Mixture of Gaussians

- When modeling real-world data, Gaussian assumption may not be appropriate.

- Consider the following example: Old Faithful Dataset
Mixture of Gaussians

• We can combine simple models into a complex model by defining a superposition of $K$ Gaussian densities of the form:

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

• Note that each Gaussian component has its own mean $\mu_k$ and covariance $\Sigma_k$. The parameters $\pi_k$ are called mixing coefficients.

• More generally, mixture models can comprise linear combinations of other distributions.

\[ \forall k : \pi_k \geq 0 \quad \sum_{k=1}^{K} \pi_k = 1 \]
Mixture of Gaussians

- Illustration of a mixture of 3 Gaussians in a 2-dimensional space:

(a) Contours of constant density of each of the mixture components, along with the mixing coefficients

(b) Contours of marginal probability density \( p(x) = \sum_{k=1}^{K} \pi_k N(x | \mu_k, \Sigma_k) \)

(c) A surface plot of the distribution \( p(x) \).
Maximum Likelihood Estimation

• Given a dataset $D$, we can determine model parameters $\mu_k, \Sigma_k, \pi_k$ by maximizing the log-likelihood function:

$$\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\}$$

Log of a sum: no closed form solution

• **Solution**: use standard, iterative, numeric optimization methods or the Expectation Maximization algorithm.
The Exponential Family

- The exponential family of distributions over \( \mathbf{x} \) is defined to be a set of distributions of the form:

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

where

- \( \eta \) is the vector of natural parameters
- \( u(x) \) is the vector of sufficient statistics

- The function \( g(\eta) \) can be interpreted as the coefficient that ensures that the distribution \( p(x|\eta) \) is normalized:

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

• The Bernoulli distribution is a member of the exponential family:

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

• Comparing with the general form of the exponential family:

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

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
Bernoulli Distribution

• The Bernoulli distribution is a member of the exponential family:

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

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

• The Bernoulli distribution can therefore be written as:

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

where

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

• The Multinomial distribution is a member of the exponential family:

\[ p(\mathbf{x} | \mathbf{\mu}) = \prod_{k=1}^{M} \mu_{k}^{x_{k}} = \exp \left\{ \sum_{k=1}^{M} x_{k} \ln \mu_{k} \right\} = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \left( \boldsymbol{\eta}^{T} \mathbf{u}(\mathbf{x}) \right) \]

where \( \mathbf{x} = (x_{1}, \ldots, x_{M})^{T} \) \( \boldsymbol{\eta} = (\eta_{1}, \ldots, \eta_{M})^{T} \)

and

\[ \eta_{k} = \ln \mu_{k} \]
\[ \mathbf{u}(\mathbf{x}) = \mathbf{x} \]
\[ h(\mathbf{x}) = 1 \]
\[ g(\boldsymbol{\eta}) = 1. \]

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

• In some cases it will be convenient to remove the constraint by expressing the distribution over the M-1 parameters.
Multinomial Distribution

• The Multinomial distribution is a member of the exponential family:

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

• Let \( \mu_M = 1 - \sum_{k=1}^{M-1} \mu_k \)

• This leads to:

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

• Here the parameters \( \eta_k \) are independent.

• Note that:

\[ 0 \leq \mu_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{M-1} \mu_k \leq 1. \]
Multinomial Distribution

- The Multinomial distribution is a member of the exponential family:

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

- The Multinomial distribution can therefore be written as:

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

where

\[ \eta = (\eta_1, \ldots, \eta_{M-1}, 0)^T \]

\[ u(x) = x \]

\[ h(x) = 1 \]

\[ g(\eta) = \left( 1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right)^{-1} \]
Gaussian Distribution

• The Gaussian distribution can be written as:

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

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

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

where

\[
\eta = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \quad h(x) = (2\pi)^{-1/2}
\]

\[
u(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad g(\eta) = (-2\eta_2)^{1/2} \exp \left( \frac{\eta_1^2}{4\eta_2} \right).
\]
ML for the Exponential Family

- Remember the Exponential Family:
  \[ p(x|\eta) = h(x)g(\eta) \exp \left\{ \eta^T u(x) \right\} \]

- From the definition of the normalizer \( g(\eta) \):
  \[ g(\eta) \int h(x) \exp \left\{ \eta^T u(x) \right\} \, dx = 1 \]

- We can take a derivative w.r.t \( \eta \):
  \[ \nabla g(\eta) \int h(x) \exp \left\{ \eta^T u(x) \right\} \, dx + g(\eta) \int h(x) \exp \left\{ \eta^T u(x) \right\} u(x) \, dx = 0 \]

- Thus
  \[ -\nabla \ln g(\eta) = \mathbb{E}[u(x)] \]
ML for the Exponential Family

- Remember the Exponential Family:
  \[ p(x|\eta) = h(x)g(\eta) \exp \{ \eta^T u(x) \} \]

- We can take a derivative w.r.t \( \eta \):
  \[
  \nabla g(\eta) \int h(x) \exp \{ \eta^T u(x) \} \, dx + g(\eta) \int h(x) \exp \{ \eta^T u(x) \} \, u(x) \, dx = 0
  \]

  \[
  1/g(\eta) \quad \text{and} \quad \mathbb{E}[u(x)]
  \]

- Thus
  \[
  -\nabla \ln g(\eta) = \mathbb{E}[u(x)]
  \]

- Note that the covariance of \( u(x) \) can be expressed in terms of the second derivative of \( g(\eta) \), and similarly for the higher moments.
ML for the Exponential Family

• Suppose we observed i.i.d data \( X = \{x_1, \ldots, x_N\} \).

• We can construct the log-likelihood function, which is a function of the natural parameter \( \eta \).

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

\[
p(X|\eta) = \left( \prod_{n=1}^{N} h(x_n) \right) g(\eta)^N \exp \left\{ \eta^T \sum_{n=1}^{N} u(x_n) \right\}.
\]

• Therefore we have

\[
-\nabla \ln g(\eta_{ML}) = \frac{1}{N} \sum_{n=1}^{N} u(x_n)
\]

Sufficient Statistic