Deep Learning - Foundations and Concepts

Chapter 3. Standard Distributions

nonlineark@github

February 8, 2025

Outline

1. Discrete Variables
2. The Multivariate Gaussian
3. Periodic Variables
4. The Exponential Family
5. Nonparametric Methods

Bernoulli distribution

- Consider a binary random variable $x \in \{0, 1\}$ and a parameter $0 \leq \mu \leq 1$, such that $p(x = 1) = \mu$ and $p(x = 0) = 1 - \mu$.
- Probability distribution: $\mathrm{Bern}(x; \mu) = \mu^x (1 - \mu)^{1-x}$.
- Expectation: $E(x) = \mu$.
- Variance: $\mathrm{var}(x) = \mu(1 - \mu)$.

Bernoulli distribution

Model the Bernoulli distribution given observations $\{x_1, \ldots, x_N\}$.

$$
\begin{aligned}
p(x_1, \ldots, x_N; \mu) &= \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n} \\
\log p(x_1, \ldots, x_N; \mu) &= \sum_{n=1}^{N} (x_n \log \mu + (1-x_n) \log(1-\mu)) \\
&= \log \mu \sum_{n=1}^{N} x_n + \log(1-\mu)(N - \sum_{n=1}^{N} x_n) \\
\mu_{ML} &= \frac{1}{N} \sum_{n=1}^{N} x_n
\end{aligned}
$$

Binomial distribution

- Consider a random variable $m = \sum_{n=1}^{N} x_n$, where x_n are independent random variables obey Bernoulli distribution with parameter μ.
- Probability distribution: $\mathrm{Bin}(m; N, \mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$.
- Expectation: $E(m) = N\mu$.
- Variance: $\mathrm{var}(m) = N\mu(1-\mu)$.

Multinomial distribution

- Consider a random variable $x \in \{\mathrm{e}_1, \ldots, \mathrm{e}_K\}$ and a parameter $\mu \in \mathbb{R}^K$, such that $p(x = \mathrm{e}_k) = \mu_k$.
- Probability distribution: $p(x; \mu) = \prod_{k=1}^{K} \mu_k^{x_k}$.
- Expectation: $E(x) = \mu$.
- Covariance: $\mathrm{cov}(x) = \mathrm{diag}(\mu_1, \ldots, \mu_K) - \mu\mu^T$.

Multinomial distribution

Model the generalized Bernoulli distribution given observations $x^1, \ldots, x^N$.

$$p(x^1, \ldots, x^N; \mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_k^n}$$

$$\log p(x^1, \ldots, x^N; \mu) = \sum_{n=1}^{N} \sum_{k=1}^{K} x_k^n \log \mu_k = \sum_{k=1}^{K} (\sum_{n=1}^{N} x_k^n) \log \mu_k$$

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

For the last step, we used Lagrange multiplier to take into the constraint $\sum_{k=1}^{K} \mu_k = 1$.

Multinomial distribution

- Consider a random variable $m = \sum_{n=1}^{N} x^n$, where x^n are independent random variables obey the generalized Bernoulli distribution with parameter μ.
- Probability distribution: $\mathrm{Mult}(m; N, \mu) = \frac{N!}{\prod_{k=1}^{K} m_k!} \prod_{k=1}^{K} \mu_k^{m_k}$.
- Expectation: $E(m) = N\mu$.
- Covariance: $\mathrm{cov}(m) = N(\mathrm{diag}(\mu_1, \ldots, \mu_K) - \mu\mu^T)$.

Definition

For a single variable x, the Gaussian distribution can be written in the form:

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

where μ is the mean and σ^2 is the variance. For a D-dimensional vector x, the multivariate Gaussian distribution takes the form:

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}(\det\Sigma)^{\frac{1}{2}}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

where μ is the D-dimensional mean vector, Σ is the $D \times D$ covariance matrix.

Geometry of the Gaussian

Without loss of generality, we assume Σ is symmetric. As a self-adjoint operator, there exists an orthonormal basis $(u_1, \ldots, u_D)$ under which Σ is diagonalized:

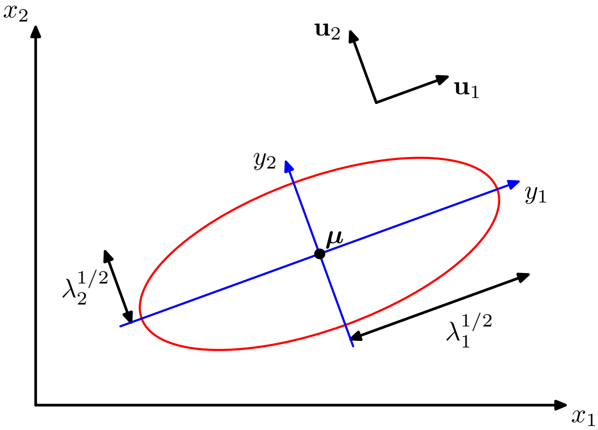
$$\operatorname{diag}(\lambda_1, \ldots, \lambda_D) = U^T \Sigma U$$

where U is the orthogonal matrix whose jth column is u_j. Now let $x - \mu = Uy$, we see that under the new basis, the multivariate Gaussian takes the form:

$$\begin{aligned}
\mathcal{N}(x; \mu, \Sigma) &= \frac{1}{(2\pi)^{\frac{D}{2}} (\lambda_1 \ldots \lambda_D)^{\frac{1}{2}}} \exp(-\frac{1}{2} y^T \operatorname{diag}^{-1}(\lambda_1, \ldots, \lambda_D) y) |\det U| \\
&= \frac{1}{\sqrt{2\pi\lambda_1} \ldots \sqrt{2\pi\lambda_D}} \exp(-\frac{1}{2} \sum_{d=1}^{D} \frac{y_d^2}{\lambda_d}) \\
&= \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d}} \exp(-\frac{y_d^2}{2\lambda_d})
\end{aligned}$$

Geometry of the Gaussian

Figure: Geometry of the Gaussian



Geometry of the Gaussian

It's easy to see that the multivariate Gaussian is indeed normalized:

$$\begin{aligned}\int \mathcal{N}(x;\mu,\Sigma)\mathrm{d}x &= \int \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d}} \exp(-\frac{y_d^2}{2\lambda_d})\mathrm{d}y \\ &= \prod_{d=1}^{D} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\lambda_d}} \exp(-\frac{y_d^2}{2\lambda_d})\mathrm{d}y_d \\ &= 1\end{aligned}$$

Expectation and covariance

Similarly, we can calculate the expectation and covariance of the multivariate Gaussian:

$$
\begin{aligned}
E(x) &= \int \mathcal{N}(x;\mu,\Sigma)x\mathrm{d}x \\
&= \int \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d}} \exp(-\frac{y_d^2}{2\lambda_d})(\mu + Uy)|\det U|\mathrm{d}y \\
&= \mu + U \int \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d}} \exp(-\frac{y_d^2}{2\lambda_d}) y\mathrm{d}y \\
&= \mu
\end{aligned}
$$

Expectation and covariance

$$
\begin{aligned}
E(xx^T) &= \int \mathcal{N}(x; \mu, \Sigma) xx^T \mathrm{d}x \\
&= \int \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d}} \exp(-\frac{y_d^2}{2\lambda_d})(\mu + Uy)(\mu + Uy)^T |\det U| \mathrm{d}y \\
&= \mu\mu^T + U(\int \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d}} \exp(-\frac{y_d^2}{2\lambda_d}) yy^T \mathrm{d}y)U^T \\
&= \mu\mu^T + U \mathrm{diag}(\lambda_1, \ldots, \lambda_D) U^T = \mu\mu^T + \Sigma \\
\mathrm{cov}(x) &= E(xx^T) - E(x)E(x^T) = \Sigma
\end{aligned}
$$

The good and the bad about the Gaussian

- The Gaussian distribution arises in many different contexts:
  - The distribution that maximizes the entropy is the Gaussian.
  - Central limit theorem.
- The Gaussian distribution has many important analytical properties.
- For large D, the total number of parameters grows quadratically with D, manipulating and inverting the large matrices can become prohibitive.
- The Gaussian distribution is intrinsically unimodal, and so is unable to provide a good approximation to multimodal distributions.

Conditional distribution and marginal distribution

Problem

Suppose x obeys the Gaussian distribution $\mathcal{N}(x; \mu, \Sigma)$. If we partition x into x_a and x_b, that is $x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$, what are the expressions for the conditional distribution $p(x_a|x_b)$ and the marginal distribution $p(x_a)$?

Conditional distribution and marginal distribution

First step, let's also partition the mean and covariance accordingly:

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

Because Σ^{-1} (called the precision matrix) appears frequently, we also partition Σ^{-1}:

$$\Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

Notice that because Σ and Σ^{-1} are symmetric, Σ_{aa}, Σ_{bb}, Λ_{aa} and Λ_{bb} are symmetric as well. Further, we have $\Sigma_{ba} = \Sigma_{ab}^T$ and $\Lambda_{ba} = \Lambda_{ab}^T$.

Conditional distribution and marginal distribution

Second step, let's complete the square! Notice:

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = x\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu$$

For conditional distribution $p(x_a|x_b)$:

$$\begin{aligned}(x-\mu)^T\Sigma^{-1}(x-\mu) &= \begin{pmatrix} x_a^T - \mu_a^T & x_b^T - \mu_b^T \end{pmatrix} \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \begin{pmatrix} x_a - \mu_a \\ x_b - \mu_b \end{pmatrix} \\ &= x_a^T\Lambda_{aa}x_a - 2x_a^T(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)) + \text{const}\end{aligned}$$

Compare, we see:

$$\begin{aligned}\Sigma_{x_a|x_b}^{-1} &= \Lambda_{aa} \\ \Sigma_{x_a|x_b} &= \Lambda_{aa}^{-1} \\ \Sigma_{x_a|x_b}^{-1}\mu_{x_a|x_b} &= \Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b) \\ \mu_{x_a|x_b} &= \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)\end{aligned}$$

Conditional distribution and marginal distribution

For marginal distribution, $p(x_a) = \int p(x_a, x_b)\mathrm{d}x_b$. Let's first complete the square for x_b to integrate it out:

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \begin{pmatrix} x_a^T - \mu_a^T & x_b^T - \mu_b^T \end{pmatrix} \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \begin{pmatrix} x_a - \mu_a \\ x_b - \mu_b \end{pmatrix}$$

$$= x_b^T \Lambda_{bb} x_b - 2x_b^T \Lambda_{bb}(\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)) + \cdots$$

$$= x_b^T \Lambda_{bb} x_b - 2x_b^T \Lambda_{bb} m + m^T \Lambda_{bb} m + \cdots$$

$$= (x_b - m)^T \Lambda_{bb}(x_b - m) + \cdots$$

where $m = \mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)$. We see that when integrating x_b, the result will be a constant not depending on x_a, although m depends on x_a.

Conditional distribution and marginal distribution

Which means, we can take a look at the terms left in $\cdots$, and complete the square for x_a to get the mean and the covariance for x_a:

$$\cdots = (x_a - \mu_a)^T (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})(x_a - \mu_a)$$

We see that:

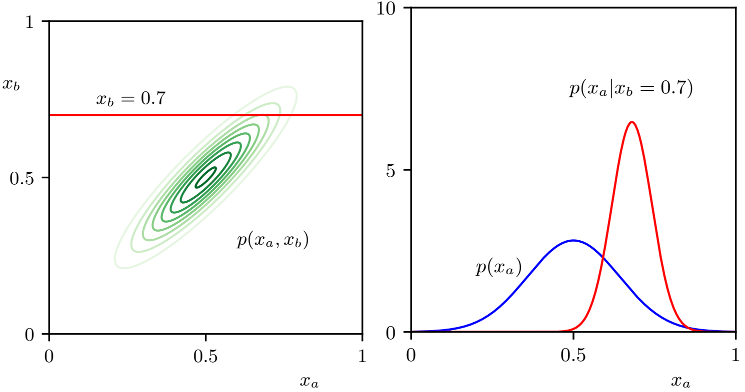
$$\begin{aligned}
\Sigma_{x_a} &= (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1} \\
\mu_{x_a} &= \mu_a
\end{aligned}$$

Through a rather ugly equation (known as Schur complement), we can simplify the expression for Σ_{x_a} to a much nicer one:

$$\Sigma_{x_a} = \Sigma_{aa}$$

Conditional distribution and marginal distribution

Figure: The marginal distribution and the conditional distribution



Bayes' theorem

Problem

Suppose that we are given a Gaussian marginal distribution $p(x)$ and a Gaussian conditional distribution $p(y|x)$. What are the expressions for the marginal distribution $p(y)$ and the conditional distribution $p(x|y)$?

Bayes' theorem

To make things easier, we suppose that $p(y|x)$ has a mean that is a linear function of x and a covariance that is independent of x:

$$p(x) = \mathcal{N}(x; \mu, \Lambda^{-1})$$
$$p(y|x) = \mathcal{N}(y; Ax + b, L^{-1})$$

Bayes' theorem

Let's find the joint distribution of $p(x, y)$, then from $p(x, y)$ we can easily get both $p(y)$ and $p(x|y)$:

$$
\begin{aligned}
&(x-\mu)^T \Lambda (x-\mu) + (y-(Ax+b))^T L (y-(Ax+b)) \\
&= \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - 2 \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Lambda\mu - A^T L b \\ Lb \end{pmatrix}
\end{aligned}
$$

Using the (ugly but useful) Schur complement again, we have:

$$
\begin{aligned}
\Lambda_{x,y} &= \begin{pmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{pmatrix} \\
\Sigma_{x,y} = \Lambda_{x,y}^{-1} &= \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1} A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1} A^T \end{pmatrix} \\
\mu_{x,y} = \Sigma_{x,y} \begin{pmatrix} \Lambda\mu - A^T L b \\ Lb \end{pmatrix} &= \begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}
\end{aligned}
$$

Bayes' theorem

From the joint distribution of $p(x, y)$, we can easily get:

$$
\begin{aligned}
\Sigma_y &= L^{-1} + A\Lambda^{-1}A^T \\
\mu_y &= A\mu + b \\
\Lambda_{x|y} &= \Lambda + A^T LA \\
\mu_{x|y} &= \mu - (\Lambda + A^T LA)^{-1}(-A^T L)(y - (A\mu + b)) \\
&= (\Lambda + A^T LA)^{-1}(A^T L(y - b) + \Lambda\mu)
\end{aligned}
$$

Maximum likelihood

Problem

We have N observations of a random variable x: $x_1, \ldots, x_N$ that are drawn independently from a multivariate Gaussian distribution whose mean μ and covariance Σ are unknown. How do we determine these parameters from the data set?

Maximum likelihood

$$L = -\log p(x_1, \ldots, x_N; \mu, \Lambda^{-1})$$
$$= \frac{ND}{2}\log(2\pi) - \frac{N}{2}\log\det\Lambda + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^T\Lambda(x_n - \mu)$$

$$\frac{\partial L}{\partial \mu} = N(\mu - \frac{1}{N}\sum_{n=1}^{N} x_n)^T\Lambda \qquad \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{\partial L}{\partial \Lambda}(\Lambda)H = \frac{N}{2}\mathrm{tr}((\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)(x_n - \mu)^T - \Lambda^{-1})H)$$

$$\Sigma_{ML} = \Lambda_{ML}^{-1} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T$$

Maximum likelihood

A couple of more words regarding $\frac{\partial L}{\partial \Lambda}$. The only thing that needs more explanation is how to differentiate $\log \det X$:

$$\lim_{h\to 0} \frac{1}{h}(\det(X + he_{ij}) - \det X) = \lim_{h\to 0} \frac{1}{h}(\det X + hX_{ij} - \det X) = X_{ij}$$

where X_{ij} is the ij-cofactor of X. Now we can calculate $D\det$ easily:

$$D\det(X)H = \sum_{i,j} X_{ij}h_{ij} = \mathrm{tr}((\mathrm{cof}(X))^T H)$$

where $\mathrm{cof}(X)$ is the cofactor matrix of X. From here we have:

$$D\log\det(X)H = \frac{1}{\det X}\mathrm{tr}((\mathrm{cof}\, X)^T H) = \mathrm{tr}(X^{-1}H)$$

Maximum likelihood

Similarly to univariate Gaussian, we find that Σ_{ML} is biased:

$$E(\mu_{ML}) = \mu$$
$$E(\Sigma_{ML}) = \frac{N-1}{N}\Sigma$$

We can correct this bias by defining a different estimator $\tilde{\Sigma}$ given by:

$$\tilde{\Sigma} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T$$

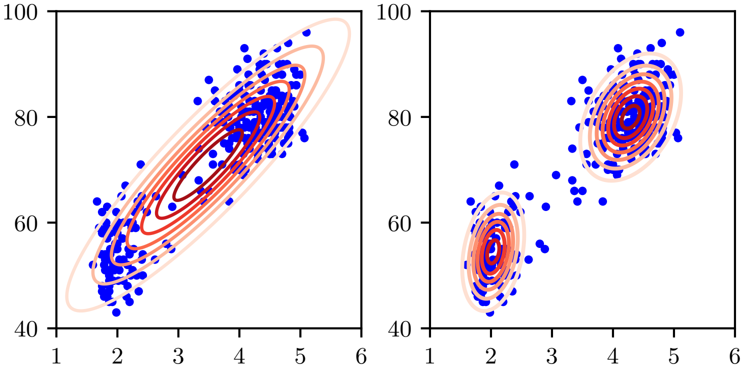
Sequential estimation

Because μ_{ML} only depends on the sum of the data points, it allows us to process the data points one at a time. If we denote by μ_{ML}^{N} the result for the maximum likelihood estimator of the mean when it is based on N observations:

$$
\begin{aligned}
\mu_{ML}^{N} &= \frac{1}{N}\sum_{n=1}^{N} x_n \\
&= \frac{1}{N}((N-1)\mu_{ML}^{N-1} + x_N) \\
&= \mu_{ML}^{N-1} + \frac{1}{N}(x_N - \mu_{ML}^{N-1})
\end{aligned}
$$

Mixtures of Gaussians

Figure: A single Gaussian fails to capture the two clumps while a linear combination of two Gaussians gives a better representation



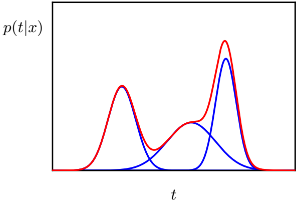
Mixtures of Gaussians

A mixture of Gaussians is a superposition of K Gaussian densities:

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

where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$.

Figure: Example of a Gaussian mixture distribution



Periodic variables

Problem

Evaluating the mean of a set of observations $\{\theta_1, \dots, \theta_N\}$ of a periodic variable θ where θ is measured in radians.

Periodic variables

Consider this as a 2-dimensional problem instead of a 1-dimensional one. Each θ_n corresponds to a point x_n on the unit circle, let's find the angle $\bar{\theta}$ for the average of these points $\bar{x}$:

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\sum_{n=1}^{N}\begin{pmatrix}\cos\theta_n \\ \sin\theta_n\end{pmatrix} = \begin{pmatrix}\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n \\ \frac{1}{N}\sum_{n=1}^{N}\sin\theta_n\end{pmatrix}$$

$$\tan\bar{\theta} = \frac{\sum_{n=1}^{N}\sin\theta_n}{\sum_{n=1}^{N}\cos\theta_n}$$

Von Mises distribution

Periodic probability density:

$$p(\theta) \geq 0$$

$$\int_0^{2\pi} p(\theta)\mathrm{d}\theta = 1$$

$$p(\theta + 2\pi) = p(\theta)$$

Is there a periodic probability density $p(\theta)$ that gives the result $\tan\bar{\theta} = \frac{\sum_{n=1}^{N}\sin\theta_n}{\sum_{n=1}^{N}\cos\theta_n}$ as a maximum likelihood estimator?

Von Mises distribution

Let's consider a 2-dimensional Gaussian conditioning on the unit circle, where the mean $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = r_0 \begin{pmatrix} \cos\theta_0 \\ \sin\theta_0 \end{pmatrix}$ and the covariance $\Sigma = \sigma^2 I$:

$$
\begin{aligned}
p(x_1, x_2) &= \frac{1}{2\pi\sigma^2} \exp(-\frac{(x_1-\mu_1)^2 + (x_2-\mu_2)^2}{2\sigma^2}) \\
p(r, \theta) &= \frac{1}{2\pi\sigma^2} \exp(-\frac{r^2 - 2r_0 r\cos(\theta-\theta_0) + r_0^2}{2\sigma^2}) r \\
p(\theta | r = 1) &= C \exp(\frac{r_0}{\sigma^2}\cos(\theta-\theta_0))
\end{aligned}
$$

Let $m = \frac{r_0}{\sigma^2}$, and normalize the constant C, we have:

$$
p(\theta; \theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m\cos(\theta-\theta_0))
$$

Von Mises distribution

Let's consider the maximum likelihood estimator for the parameter θ_0:

$$\begin{aligned} L &= \log p(\theta_1, \ldots, \theta_N; \theta_0, m) \\ &= -N \log(2\pi I_0(m)) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0) \end{aligned}$$

$$\frac{\partial L}{\partial \theta_0} = m \sum_{n=1}^{N} \sin(\theta_n - \theta_0) = m(\cos\theta_0 \sum_{n=1}^{N} \sin\theta_n - \sin\theta_0 \sum_{n=1}^{N} \cos\theta_n)$$

We indeed have:

$$\theta_0^{ML} = \frac{\sum_{n=1}^{N} \sin\theta_n}{\sum_{n=1}^{N} \cos\theta_n}$$

The exponential family

The exponential family of distributions over x, given parameters η, is defined to be the set of distributions of the form:

$$p(x;\eta) = h(x)g(\eta)\exp(\eta^T u(x))$$

The exponential family

The Bernoulli distribution is a member of the exponential family:

$$
\begin{aligned}
\mathrm{Bern}(x;\mu) &= \mu^x(1-\mu)^{1-x} \\
&= (1-\mu)\exp(x\log\frac{\mu}{1-\mu}) \\
\eta &= \log\frac{\mu}{1-\mu} \\
g(\eta) &= \frac{1}{1+\exp(\eta)} \\
u(x) &= x \\
h(x) &= 1
\end{aligned}
$$

The exponential family

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

$$
\begin{aligned}
p(x;\mu) &= \prod_{k=1}^{K} \mu_k^{x_k} = \exp(\sum_{k=1}^{K} x_k \log \mu_k) \\
&= \exp(\sum_{k=1}^{K-1} x_k \log \mu_k + (1 - \sum_{k=1}^{K-1} x_k) \log(1 - \sum_{k=1}^{K-1} \mu_k)) \\
&= (1 - \sum_{k=1}^{K-1} \mu_k) \exp(\sum_{k=1}^{K-1} x_k \log \frac{\mu_k}{1 - \sum_{k=1}^{K-1} \mu_k})
\end{aligned}
$$

The exponential family

By comparison to the standard form, we have:

$$\eta_k = \log \frac{\mu_k}{1 - \sum_{k=1}^{K-1} \mu_k}$$

$$g(\eta) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\eta_k)}$$

$$u(x) = x$$

$$h(x) = 1$$

The exponential family

The Gaussian distribution is a member of the exponential family:

$$
\begin{aligned}
\mathcal{N}(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{\mu^2}{2\sigma^2})\exp(\begin{pmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{pmatrix}^T\begin{pmatrix}x\\ x^2\end{pmatrix})
\end{aligned}
$$

The exponential family

By comparison to the standard form, we have:

$$\eta = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}$$

$$g(\eta) = \sqrt{-\frac{\eta_2}{\pi}} \exp(\frac{\eta_1^2}{4\eta_2})$$

$$u(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

$$h(x) = 1$$

Sufficient statistics

Let's estimate the parameter η in the exponential family distribution using the technique of maximum likelihood:

$$
\begin{aligned}
L &= \log p(x_1, \ldots, x_N; \eta) \\
&= \sum_{n=1}^{N} \log(h(x_n)) + N \log(g(\eta)) + \eta^T \sum_{n=1}^{N} u(x_n) \\
\frac{\partial L}{\partial \eta} &= N D \log(g(\eta)) + \sum_{n=1}^{N} (u(x_n))^T \\
-(D \log(g(\eta_{ML})))^T &= \frac{1}{N} \sum_{n=1}^{N} u(x_n)
\end{aligned}
$$

Sufficient statistics

We see the estimation is plausible by looking at the expectation of $u(x)$:

$$g(\eta)\int h(x)\exp(\eta^T u(x))\mathrm{d}x = 1$$

$$Dg(\eta)\int h(x)\exp(\eta^T u(x))\mathrm{d}x + g(\eta)\int h(x)\exp(\eta^T u(x))(u(x))^T\mathrm{d}x = 0$$

$$\frac{Dg(\eta)}{g(\eta)} + E(u(x)^T) = 0$$

$$E(u(x)) = -(D\log(g(\eta)))^T$$

Histograms

- Partition x into distinct bins of width Δ_i.
- Count the number n_i of observations of x falling in bin i.
- $p_i = \frac{n_i}{N\Delta_i}$.
- Δ_i plays the role of a smoothing parameter, best results are obtained for some intermediate value of Δ_i.
- Curse of dimensionality.

Kernel densities

To estimate the probability density of $p(x)$:

- Consider some small region $\mathcal{R}$ that contains x.
- Suppose we have collected a data set comprising of N observations, and there are K points that lie inside $\mathcal{R}$.
- And the region $\mathcal{R}$ has volume V.

Then we obtain our density estimate in the form:

$$p(x) = \frac{K}{NV}$$

We can exploit this result in two different ways:

- Fix V and determine K from the data: The kernel approach.
- Fix K and determine V from the data: The K-nearest-neighbor technique.

Kernel densities

Parzen window:

$$k(u) = \begin{cases} 1, & |u_i| \leq \frac{1}{2}, 1 \leq i \leq D \\ 0, & \text{otherwise} \end{cases}$$

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k(\frac{x - x_n}{h})$$

Kernel densities

We can choose any other kernal function $k(u)$ subject to the conditions:

$$k(u) \geq 0$$
$$\int k(u)\mathrm{d}u = 1$$

For example, we could choose the Gaussian:

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{\frac{D}{2}}} \exp(-\frac{||x - x_n||^2}{2h^2})$$

where h plays the role of a smoothing parameter.

Nearest-neighbors

To estimate the probability density of $p(x)$:

- Consider a small sphere centered on the point x.
- Allow the radius of the sphere to grow until it contains precisely K data points.
- K plays the role of a smoothing parameter.

The K-nearest-neighbor technique can be extended to the problem of classification.