Deep Learning - Foundations and Concepts

Chapter 3. Standard Distributions

nonlineark@github

February 7, 2025

Outline

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 $\{x_1, \ldots, x_N\}$.

$$p(x_1, \ldots, x_N; \mu) = \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n}$$

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

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

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 $x^1, \dots, x^N$.

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

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

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

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:

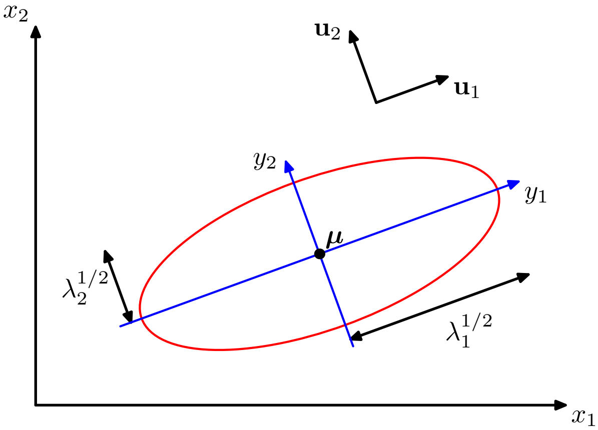
$$\mathrm{diag}(\lambda_1, \ldots, \lambda_D) = U^T \Sigma U$$

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:

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

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 is the expression 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:

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

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

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