Probability and Distributions

Mathematics for Machine Learning

San Diego Machine Learning
Liam Barstad

Distributions

Probability distribution – equation that describes likelihood of random variable(s)

Sample space (Ω) – set of all possible outcomes –
{HH, HT, TH, TT}

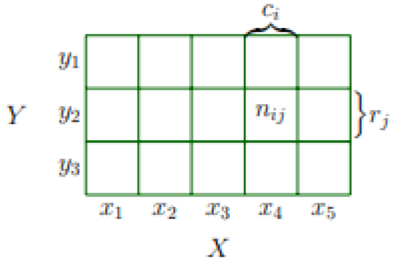
Event space (*A*) – subset of sample space w/ outcome –
(e.g. at least one head) – {HH, HT, TH}

Target space (*T*) – set of values that can result -
(e.g. number of heads) – {0, 1, 2}

Random Variable (X) – mapping where $X : \Omega \rightarrow T$

Discrete Distribution Functions

Probability mass function (pmf) – describes discrete random variable – p(x, y)



$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N},$$

Marginal probability p(x), p(y)

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{3} n_{ij}}{N}$$

$$P(Y = y_j) = \frac{r_j}{N} = \frac{\sum_{i=1}^{5} n_{ij}}{N},$$

Conditional probability

p(y | x) – $P(Y = y_j \,|\, X = x_i) = \dfrac{n_{ij}}{c_i},$

p(x | y) – $P(X = x_i \,|\, Y = y_j) = \dfrac{n_{ij}}{r_j}.$

Continuous Distribution Functions

Probability density function (pdf) – describes continuous random variable

Definition 6.1 (Probability Density Function). A function $f : \mathbb{R}^D \to \mathbb{R}$ is called a *probability density function* (*pdf*) if

1. $\forall \boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) \geqslant 0$
2. Its integral exists and

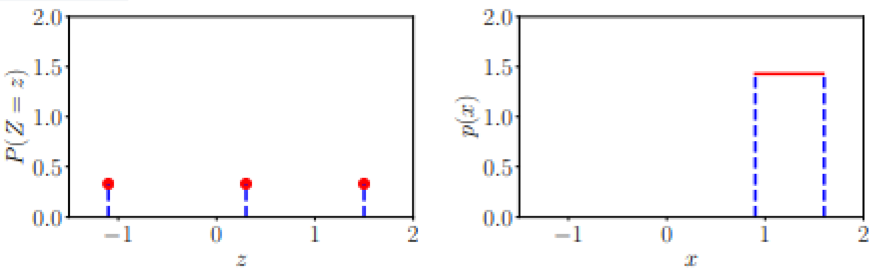
$$\int_{\mathbb{R}^D} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1 \,. \tag{6.15}$$

$$P(a \leqslant X \leqslant b) = \int_a^b f(x) \mathrm{d}x \,, \tag{6.16}$$

Cumulative distribution function (cdf) – describes likelihood random variable(s) are less t

$$F_X(\boldsymbol{x}) = P(X_1 \leqslant x_1, \ldots, X_D \leqslant x_D) \,,$$

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \ldots, z_D) \mathrm{d}z_1 \cdots \mathrm{d}z_D \,.$$



(a) Discrete distribution (b) Continuous distribution

Uniform distribution – all results are

Probability Rules

Sum rule, aka marginalization property

$$p(\boldsymbol{x}) = \begin{cases} \sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) & \text{if } \boldsymbol{y} \text{ is discrete} \\ \int_{\mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} & \text{if } \boldsymbol{y} \text{ is continuous} \end{cases}$$

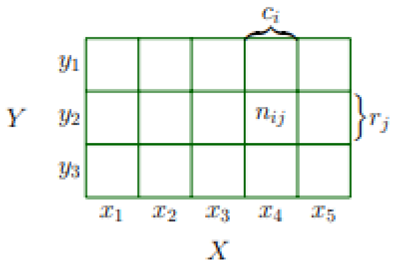
$$p(x_i) = \int p(x_1, \ldots, x_D) \mathrm{d}\boldsymbol{x}_{\setminus i}$$

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x})$$

Product rule

Bayes rule

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$



Expected Value

Each value multiplied by its probability

$$\mathbb{E}_{x_d}[x_d] := \begin{cases} \int_{\mathcal{X}} x_d p(x_d) \mathrm{d}x_d \\ \sum_{x_i \in \mathcal{X}} x_i p(x_d = x_i) \end{cases}$$

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x) p(x) \mathrm{d}x$$

$$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x) p(x)$$

Mean - E[x]

$$\mathbb{E}_X[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{bmatrix} \in \mathbb{R}^D$$

$$\mathbb{E}_X[g(\boldsymbol{x})] = \begin{bmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{bmatrix} \in \mathbb{R}^D$$

$$\begin{aligned} \mathbb{E}_X[f(\boldsymbol{x})] &= \int f(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \int [a g(\boldsymbol{x}) + b h(\boldsymbol{x})] p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= a \int g(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}x + b \int h(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= a \mathbb{E}_X[g(\boldsymbol{x})] + b \mathbb{E}_X[h(\boldsymbol{x})] \,. \end{aligned}$$

Variance + Covariance

Covariance – deviation from central value of 2 random variables

$$\mathrm{Cov}_{X,Y}[x,y] := \mathbb{E}_{X,Y}\left[(x-\mathbb{E}_X[x])(y-\mathbb{E}_Y[y])\right]$$

$$\mathrm{Cov}[x,y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

$$\mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^\top = \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}]^\top \in \mathbb{R}^{D\times E}\,.$$

Correlation – normalized covariance

$$\mathrm{corr}[x,y] = \frac{\mathrm{Cov}[x,y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \in [-1,1]$$

Variance – squared deviation from E[x], covariance of random variable w/ itself

$$\begin{aligned}\mathbb{V}_X[\boldsymbol{x}] &= \mathrm{Cov}_X[\boldsymbol{x},\boldsymbol{x}]\\ &= \mathbb{E}_X[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_X[\boldsymbol{x}]\mathbb{E}_X[\boldsymbol{x}]^\top\\ &= \begin{bmatrix}\mathrm{Cov}[x_1,x_1] & \mathrm{Cov}[x_1,x_2] & \dots & \mathrm{Cov}[x_1,x_D]\\ \mathrm{Cov}[x_2,x_1] & \mathrm{Cov}[x_2,x_2] & \dots & \mathrm{Cov}[x_2,x_D]\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{Cov}[x_D,x_1] & \dots & \dots & \mathrm{Cov}[x_D,x_D]\end{bmatrix}.\end{aligned}$$

$$\mathbb{V}_X[x] := \mathbb{E}_X[(x-\mu)^2]$$

$$\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$$

$$\frac{1}{N^2}\sum_{i,j=1}^{N}(x_i-x_j)^2 = 2\left[\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2\right]$$

Variance + Covariance Useful Properties

Given affine transformation y = Ax + b

$$\mathbb{E}_Y[\boldsymbol{y}] = \mathbb{E}_X[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_X[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$$

$$\mathbb{V}_Y[\boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \mathbb{V}_X[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}_X[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$$

$$\begin{aligned}
\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}]^\top \\
&= \mathbb{E}[\boldsymbol{x}]\boldsymbol{b}^\top + \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{A}^\top \\
&= \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top + \left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\boldsymbol{A}^\top \\
&\overset{(6.38\text{b})}{=} \boldsymbol{\Sigma}\boldsymbol{A}^\top,
\end{aligned}$$

where $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ is the covariance of X.

Geometric Interpretation

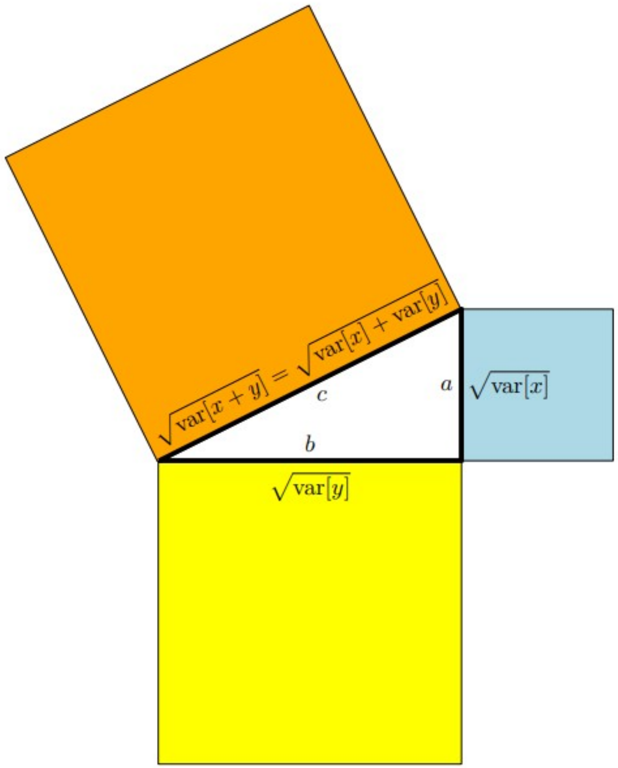
Covariance can be thought of as inner product

$$\langle X, Y \rangle := \mathrm{Cov}[x, y]$$

If inner product = 0, a and b are orthogonal, V[x + y] = V[x] + V[y]

$$\cos\theta = \frac{\langle X, Y \rangle}{\|X\| \ \|Y\|} = \frac{\mathrm{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}}$$

Information Geometry



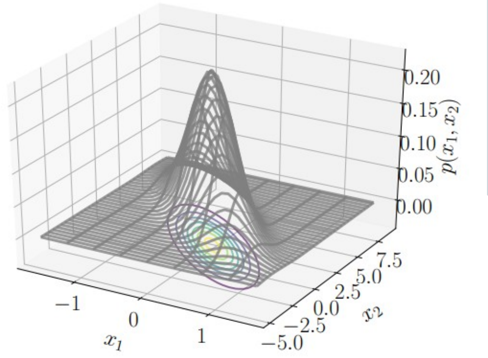
Gaussian (Normal) Distribution

Univariate density

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Multivariate density

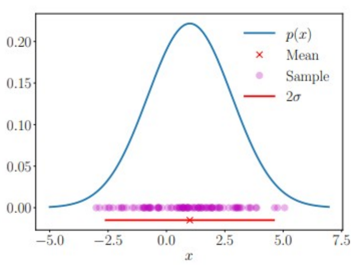
$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

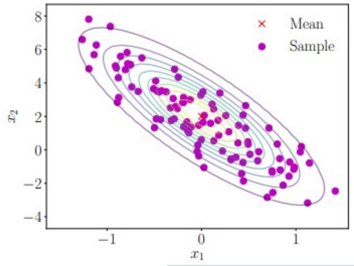


Both marginals and conditionals of gaussians are gaussian

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right)$$

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{y}) &= \mathcal{N}(\boldsymbol{\mu}_{x \mid y}, \boldsymbol{\Sigma}_{x \mid y}) \\ \boldsymbol{\mu}_{x \mid y} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \\ \boldsymbol{\Sigma}_{x \mid y} &= \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} \,. \end{aligned}$$





Properties of Gaussian Distributions

Products of Gaussians – N(x | a, A), N(x | b, B) is gaussian scaled by C, cN(x | c, C)

$$c = \mathcal{N}(\boldsymbol{a} \,|\, \boldsymbol{b},\ \boldsymbol{A} + \boldsymbol{B}) = \mathcal{N}(\boldsymbol{b} \,|\, \boldsymbol{a},\ \boldsymbol{A} + \boldsymbol{B})$$

Sums of Gaussians

$$p(\boldsymbol{x} + \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y,\ \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)$$

$$p(a\boldsymbol{x} + b\boldsymbol{y}) = \mathcal{N}(a\boldsymbol{\mu}_x + b\boldsymbol{\mu}_y,\ a^2\boldsymbol{\Sigma}_x + b^2\boldsymbol{\Sigma}_y)$$

Scalars

Common Distributions

Bernoulli Distribution – single binary random variable, represents probability of X = 1

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

Binomial Distribution – probability of observing m occurrences of X = 1 in a set of N samples

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

Beta Distribution – models uncertainty over a continuous random variable, a is number of successes, B is number of failures (i.e. there's a 95% chance the success rate is between 3% and 7%)

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1}$$

$$\Gamma(t) := \int_0^\infty x^{t-1} \exp(-x) dx, \qquad t > 0$$

$$\Gamma(t+1) = t\Gamma(t) \, .$$

Conjugacy

Prior is **conjugate** for the likelihood function if the posterior is the same form as the prior (retain same distance structure geometrically)

Example 1: Beta-Binomial, x is number of heads, μ is prob. of heads (prior)

$$p(x \mid N, \mu) = \binom{N}{x} \mu^x (1-\mu)^{N-x}$$

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

$$\begin{aligned} p(\mu \mid x = h, N, \alpha, \beta) &\propto p(x \mid N, \mu) p(\mu \mid \alpha, \beta) \\ &\propto \mu^h (1-\mu)^{(N-h)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \\ &= \mu^{h+\alpha-1}(1-\mu)^{(N-h)+\beta-1} \end{aligned}$$

Example 2: Beta-Bernoulli

$$p(\theta \mid x, \alpha, \beta) = p(x \mid \theta) p(\theta \mid \alpha, \beta)$$

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$

Exponential Family

Only family where the number of sufficient statistics used to describe data has finite dimensions

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})\right)$$

- h(x) can be absorbed into sufficient statistics by adding log(h(x)) to ϕ(x)
- A(θ) is log-partition function which makes sure the sum is 1

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})\right)$$

For a Gaussian distribution:

$$p(x \mid \boldsymbol{\theta}) \propto \exp(\theta_1 x + \theta_2 x^2)$$

$$\boldsymbol{\theta} = \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]^\top$$

$$p(x \mid \boldsymbol{\theta}) \propto \exp\left(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\right) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Exponential Family - Conjugates

Every exponential distribution has a conjugate prior

$$p(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) = h_c(\boldsymbol{\theta}) \exp\left(\left\langle \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\theta} \\ -A(\boldsymbol{\theta}) \end{bmatrix} \right\rangle - A_c(\boldsymbol{\gamma})\right)$$

For Bernoulli/Beta distribution

$$p(x \mid \mu) = \exp\left[x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right]$$

$$\boldsymbol{\gamma} := [\alpha, \beta + \alpha]^\top \text{ and } h_c(\mu) := \mu/(1-\mu)$$

$$p(\mu \mid \alpha, \beta) = \frac{\mu}{1-\mu} \exp\left[\alpha \log \frac{\mu}{1-\mu} + (\beta + \alpha)\log(1-\mu) - A_c(\boldsymbol{\gamma})\right]$$

$$p(\mu \mid \alpha, \beta) \propto \mu^{\alpha-1}(1-\mu)^{\beta-1}$$