Probability Theory

Kevin P. Murphy: Probabilistic Machine Learning Ch2 & Ch3

Presented by Wending Liu

RSE ML and Econometrics Reading Group

Oct 23 2023

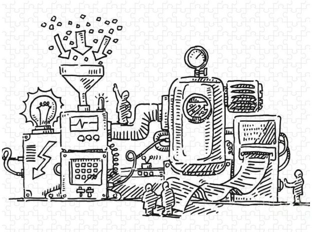
Table of Contents

Probability

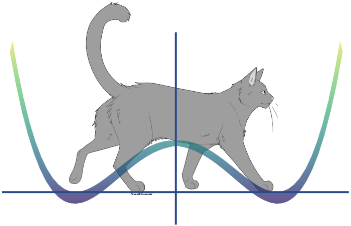
- What do we mean by saying that "the probability of a fair coin will land head is 50%"?
  - **Frequentist** interpretation: If we flip the coin 1,000,000 times, we will observe the coin land heads about 500,000 times.
  - **Bayesian** interpretation: We ***believe*** the coin is equally likely to land head or tail on our next toss.
- In the Bayesian view, probability is used to quantify our **uncertainty** about something.
- Bayesian interpretation can be used to model uncertainty about one-off events.
- Bayesian interpretation is used throughout the whole book.

Uncertainty

- But what is **uncertainty** in Bayesian interpretation?
- Epistemic/Model uncertainty
- Aleatoric/Data uncertainty



Complex machine, little human



Schrodinger's cat

Table of Contents

Bayes' rule

- Bayes' rule is a formula for computing the probability distribution over possible values of unkown quantity H given some observed data $Y = y$:

$$p(H = h \mid Y = y) = \frac{p(H = h)p(Y = y \mid H = h)}{p(Y = y)}$$

- $p(H = h)$: prior distribution.
- $p(Y = y \mid H = h)$: likelihood.
- $p(Y = y)$: marginal likelihood.
- $p(H = h \mid Y = y)$: posterior distribution.
- posterior $\propto$ prior $\times$ likelihood.

Example: Testing for Covid-19

- Suppose COVID-19 prevalence is 1% now and you take a diagnostic test that has a positive result. What's the probability that you are infected?
- H: the indicator of infection. Y: the indicator of a positive test result.
- prior: $p(H = 1) = 0.01$, $p(H = 0) = 0.99$.
- likelihood: $p(Y = 1|H = 1) = 0.875$. (true positive rate)
- false positive rate: $p(Y = 1|H = 0) = 0.025$.

$$
\begin{aligned}
& p(H = 1 \mid Y = 1) \\
= & \frac{p(Y = 1 \mid H = 1)p(H = 1)}{p(Y = 1 \mid H = 1)p(H = 1) + p(Y = 1 \mid H = 0)p(H = 0)} \\
= & \frac{\mathrm{TPR} \times \text{ prior}}{\mathrm{TPR} \times \text{ prior } + \text{ FPR } \times (1 - \text{ prior })} \\
= & \frac{0.875 \times 0.01}{0.875 \times 0.01 + 0.025 \times 0.99} = 0.261
\end{aligned}
$$

Bernoulli and binomial distribution

- $\mathrm{Ber}(y \mid \theta) \triangleq \theta^y (1-\theta)^{1-y} = \begin{cases} 1-\theta & \text{if } y = 0 \\ \theta & \text{if } y = 1 \end{cases}$
- $\mathrm{Bin}(s \mid N, \theta) \triangleq \left(\begin{array}{c} N \\ s \end{array} \right) \theta^s (1-\theta)^{N-s}$
- We want to predict a binary variable $y \in \{0, 1\}$ given some inputs $x \in \mathcal{X}$:

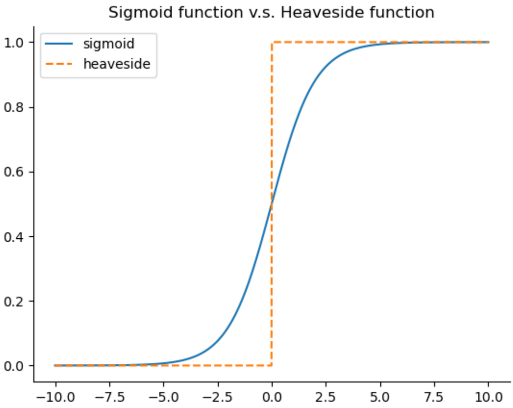
$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y \mid f(\boldsymbol{x}; \boldsymbol{\theta}))$$

- To avoid the requirement that $0 \leq f(\boldsymbol{x}; \boldsymbol{\theta}) \leq 1$, we can let f be an unconstrained function, and use the following model:

$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y \mid \sigma(f(\boldsymbol{x}; \boldsymbol{\theta})))$$

- σ is the **sigmoid** function: $\sigma(a) \triangleq \frac{1}{1+e^{-a}}$
- Binary logistic regression: $f(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{w}^T \boldsymbol{x} + b$

Sigmoid function



Categorical and multinomial distributions

- **one-hot vector**: $\boldsymbol{y} \in \{0,1\}^C$, $\sum_{c=1}^{C} y_k = 1$
- $\text{Cat}(y \mid \boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{\mathrm{I}(y=c)} = \prod_{c=1}^{C} \theta_c^{y_c}$
- $\text{Mu}(\mathbf{s} \mid N, \boldsymbol{\theta}) \triangleq \begin{pmatrix} N \\ s_1 \dots s_C \end{pmatrix} \prod_{c=1}^{C} \theta_c^{s_c}$
- $p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \text{Cat}(y \mid f(\boldsymbol{x}; \boldsymbol{\theta}))$
- We require that $0 \leq f_c(\boldsymbol{x}; \boldsymbol{\theta}) \leq 1$ and $\sum_{c=1}^{C} f_c(\boldsymbol{x}; \boldsymbol{\theta}) = 1$.
- To avoid this requirement, we pass the output from f into the **softmax function**, also called the **multinomial logit**:

$$\mathcal{S}(\boldsymbol{a}) \triangleq \left[\frac{e^{a_1}}{\sum_{c'=1}^{C} e^{a_{c'}}}, \dots, \frac{e^{a_C}}{\sum_{c'=1}^{C} e^{a_{c'}}} \right]$$

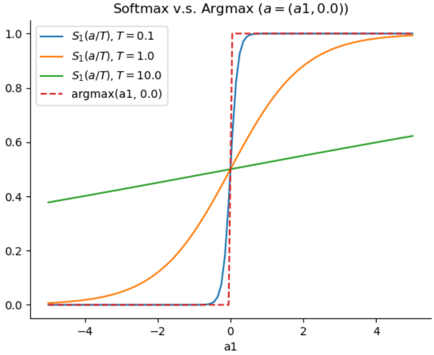
- Multinomial logistic regression: $f_c(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{w}_c^T \boldsymbol{x} + b_c$:

$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \text{Cat}(y \mid \mathcal{S}(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}))$$

Softmax function

- Softmax function is related to Boltzmann distribution in physics.
- Let us divide each a_c by a constant T called the temperature. Then as $T \to 0$, we find:

$$\mathcal{S}(\boldsymbol{a}/T)_c = \begin{cases} 1.0 & \text{if } c = \operatorname{argmax}_{c'} a_{c'} \\ 0.0 & \text{otherwise} \end{cases}$$



Guassian distribution

- Gaussian pdf: $\mathcal{N}\left(y \mid \mu, \sigma^2\right) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$
- Gaussian distribution is widely used.
  - only two parameters.
  - central limit theorem.
  - maximum entropy given fixed mean and variance.
- A robust alternative to Gaussian is the **Student's t-distribution**:

$$\mathcal{T}\left(y \mid \mu, \sigma^2, \nu\right) \propto \left[1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)}$$

  where μ is the mean, $\sigma > 0$ is the scale parameter, and $\nu > 0$ is the degree of freedom (a better term would be the **degree of normality**).
- **Laplace** distribution (heavy tail): $\mathrm{Lap}(y \mid \mu, b) \triangleq \frac{1}{2b} \exp\left(-\frac{|y-\mu|}{b}\right)$

Robustness of t distribution

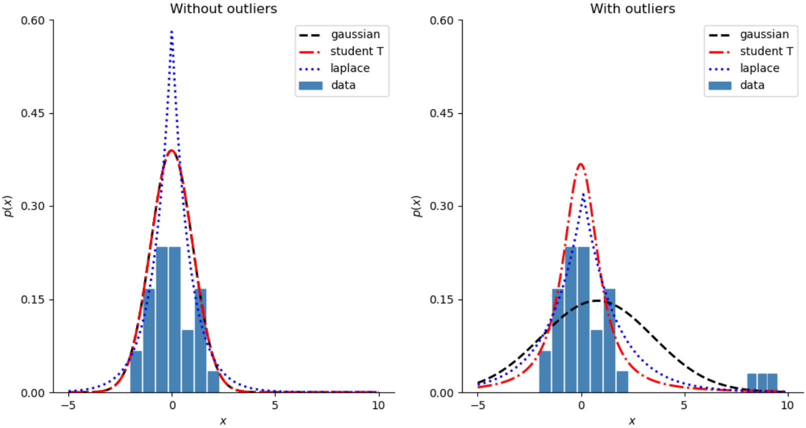
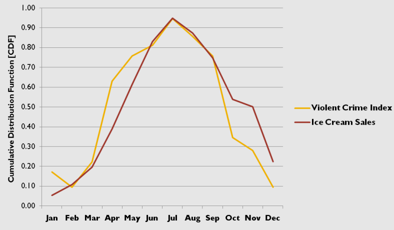


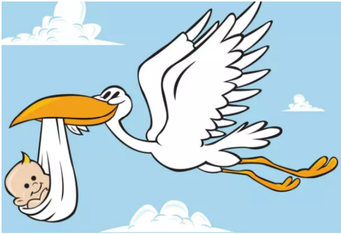
Table of Contents

Causality, Independence and Correlation

- We use multivariate models to study the dependence of variables on each other.
- causality $\not\Leftrightarrow$ dependence $\Leftarrow$ correlation



ice cream makes people angry?



storks deliver babies?

Multivariate Gaussian distribution

MVN

The MVN density is defined by the following:

$$\mathcal{N}(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]$$

where $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{y}] \in \mathbb{R}^D$ is the mean vector, and $\boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{y}]$ is the $D \times D$ covariance matrix, defined as follows:

$$\begin{aligned}
\mathrm{Cov}[\boldsymbol{y}] &\triangleq \mathbb{E}\left[(\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}])(\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}])^\top\right] \\
&= \begin{pmatrix}
\mathbb{V}[Y_1] & \mathrm{Cov}[Y_1, Y_2] & \cdots & \mathrm{Cov}[Y_1, Y_D] \\
\mathrm{Cov}[Y_2, Y_1] & \mathbb{V}[Y_2] & \cdots & \mathrm{Cov}[Y_2, Y_D] \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{Cov}[Y_D, Y_1] & \mathrm{Cov}[Y_D, Y_2] & \cdots & \mathbb{V}[Y_D]
\end{pmatrix}
\end{aligned}$$

- Suppose $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ is jointly Gaussian with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix}$$

  where $\boldsymbol{\Lambda}$ is the **precision matrix**.

Marginal Distributions

$$p(\boldsymbol{y}_1) = \mathcal{N}(\boldsymbol{y}_1 \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$
$$p(\boldsymbol{y}_2) = \mathcal{N}(\boldsymbol{y}_2 \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

Posterior Conditional Distributions

$$p(\boldsymbol{y}_1 \mid \boldsymbol{y}_2) = \mathcal{N}\left(\boldsymbol{y}_1 \mid \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right)$$
$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2}(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2)) \end{aligned}$$
$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1}$$

Linear Gussian systems

Linear Gussian system

Let $\boldsymbol{z} \in \mathbb{R}^L$ be an unknown vector of values, and $\boldsymbol{y} \in \mathbb{R}^D$ be some noisy measurement of $\boldsymbol{z}$ with the following joint distribution:

$$p(\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{z} \mid \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z\right)$$
$$p(\boldsymbol{y} \mid \boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{y} \mid \mathbf{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{\Sigma}_y\right)$$

where $\mathbf{W}$ is a matrix of size $D \times L$. The corresponding joint distribution, $p(\boldsymbol{z}, \boldsymbol{y}) = p(\boldsymbol{z})p(\boldsymbol{y} \mid \boldsymbol{z})$, is a $L + D$ dimensional Gaussian, with mean and covariance given by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_z \\ \mathbf{W}\boldsymbol{\mu}_z + \boldsymbol{b} \end{pmatrix}$$
$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z\mathbf{W}^\top \\ \mathbf{W}\boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_y + \mathbf{W}\boldsymbol{\Sigma}_z\mathbf{W}^\top \end{pmatrix}$$

Bayes rule for Gussians

$$p(\boldsymbol{z} \mid \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{z} \mid \boldsymbol{\mu}_{z|y}, \boldsymbol{\Sigma}_{z|y}\right)$$

$$\boldsymbol{\Sigma}_{z|y}^{-1} = \boldsymbol{\Sigma}_z^{-1} + \mathbf{W}^\top \boldsymbol{\Sigma}_y^{-1} \mathbf{W}$$

$$\boldsymbol{\mu}_{z|y} = \boldsymbol{\Sigma}_{z|y} \left[\mathbf{W}^\top \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_z^{-1} \boldsymbol{\mu}_z\right]$$

- Gaussian prior $p(\boldsymbol{z})$, combined with the Gaussian likelihood $p(\boldsymbol{y} \mid \boldsymbol{z})$, results in a Gaussian posterior $p(\boldsymbol{z} \mid \boldsymbol{y})$.
- Thus Gaussians are closed under Bayesian conditioning.
- The Gaussian prior is a conjugate prior for the Gaussian likelihood.

Example: Inferring an unknown scalar

- Assume we have one noisy measurement y for an unknown quantity z.
- Prior: $p(z) = \mathcal{N}(z \mid \mu_0, \Sigma_0)$.
- Likelihood: $p(y \mid z) = \mathcal{N}(y \mid z, \Sigma_y)$.
- Posterior:

$$
\begin{aligned}
p(z \mid y) &= \mathcal{N}\left(z \mid \mu_1, \Sigma_1\right) \\
\Sigma_1 &= \left(\frac{1}{\Sigma_0} + \frac{1}{\Sigma_y}\right)^{-1} = \frac{\Sigma_y \Sigma_0}{\Sigma_0 + \Sigma_y} \\
\mu_1 &= \Sigma_1 \left(\frac{\mu_0}{\Sigma_0} + \frac{y}{\Sigma_y}\right) \\
&= y - (y - \mu_0) \frac{\Sigma_y}{\Sigma_y + \Sigma_0} \quad (\textbf{shrinkage})
\end{aligned}
$$

- strong prior $\rightarrow$ large shrinkage.