Probability Tutorial

Friday 29 September 2017

Index

Basic foundations on Probability

Probability Space $(\Omega, \mathcal{F}, P)$: e.g. Throwing a dice

- Set of atomic events Ω:

$$\{1, 2, 3, 4, 5, 6\}$$

- Set of all non-atomic events $\mathcal{F} \subseteq 2^{\Omega}$

$$\text{The number is odd}$$

- Probability measure $P : \mathcal{F} \rightarrow [0, 1]$

$$P(\text{The number is odd}) = 1/2$$

$$P(\text{The number is 1}) = 1/6$$

Basic foundations on Probability

Probability Axioms

- Normalization:

$$P(\Omega) = 1$$

- Non-Negativity

$$\forall A \in \mathcal{F} : P(A) \geq 0$$

- σ-Additivity

$$\forall A_1 \dots A_n \dots \text{ s.t. } A_i \cap A_j = \emptyset \quad \forall i \neq j$$

$$P(\cup_i A_i) = \sum_i P(A_i)$$

If $A_i \cap A_j \neq \emptyset$ then the union bound holds

$$P(\cup_i A_i) \leq \sum_i P(A_i)$$

Basic foundations on Probability

Probability Rules

- Marginalization (Sum Rule):

$$f(x) = \sum_y f(x, y)$$

- Factorization (Product Rule)

$$f(x, y) = f(x|y)f(y) = f(y|x)f(x)$$

$$f(x, y, z) = f(x|y, z)f(y|z)f(z) = f(y|x, z)f(x|z)f(z)$$

- For a pdf of n variables, i.e., $f(x_1, x_2, \ldots x_n)$, how many different factorizations exist? If the variables are all independent, how many different factorizations exist?

Basic foundations on Probability

Indpendence and Conditional Independence

- x, y are independent (also, $x \perp y$) iff:

$$f(x, y) = f(x)f(y)$$

- x, y are independent given z (also, $x \perp y|z$) iff:

$$f(x, y|z) = f(x|z)f(y|z)$$

- Factorization (Product Rule) with conditional independence:

$$f(x, y, z) = f(z|x, y)f(x|y)f(y)$$

$$f(x, y, z) = f(x|y, z)f(y|z)f(z) = f(x|z)f(y|z)f(z)$$

Index

Bayes Rule

1% of women at age fourty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies.
A women in this age group had a positive mammography in a routine screeining. What is the probability that she has breast cancer?

Bayes Rule

Data

- $P(BC = T) = 1\%$, $P(BC = F) = 99\%$.
- $P(+|BC = T) = 80\%$, $P(+|BC = F) = 9.6\%$.

$$
\begin{aligned}
P(BC = T|+) &= \frac{P(BC = T, +)}{P(+)} \\
&= \frac{P(BC = T, +)}{\sum_{BC=\{T,F\}} P(BC, +)} \\
&= \frac{P(+|BC = T)P(BC = T)}{\sum_{BC=\{T,F\}} P(+|BC)P(BC)} \\
&= \frac{80\% \times 1\%}{80\% \times 1\% + 9.6\% \times 99\%} = 7.76\%
\end{aligned}
$$

Geometric Distribution

Suppose you throw a dice repeatedly until you get a 6.

(a) What is the set of atomic events Ω?

(b) What is the probability of finding a sequence of length n?

(c) What is the expected value of the sequence length?

(d) What is the expected number of 3s we observe?

Geometric Distribution

(a) Any sequence of elements that only in the last position contains a 6, e.g. $w_k = \{1, 5, 2, 3, 6\}$.

(b)

$$P(L_n) = (5/6)^{n-1}(1/6)$$

(c)

$$\begin{aligned}\mathbb{E}(L_n) &= \sum_{i=1}^{\infty} i(5/6)^{i-1}(1/6) \\ &= (1/6)\frac{1}{(1-5/6)^2} = 6\end{aligned}$$

Geometric Distribution

(d) Define event A_i : throw i is number 3.

$$P(A_i) = \underbrace{(5/6)^{i-1}}_{\text{not a 6}}\underbrace{(1/6)}_{\text{a 3}}$$

Lets call w_k the k-th sequence (e.g. $w_k = \{1,3,4,3,3,5,6\}$, $w_k \in A_2, A_4, A_5$).
Define RV S : number of 3s in outcome (e.g. $S(w_k) = 3$).
How to write S in terms of A_i?

$$S(w_k) = \sum_{i=1}^{\infty} \mathbb{1}_{A_i}(w_k)$$

$$\mathbb{1}_{A_i}(w) = \begin{cases} 1, & w \in A_i \\ 0, & \text{otherwise} \end{cases}$$

Geometric Distribution

Expected number of 3s is the expected number of S.

$$\mathbb{E}(S) = \mathbb{E}\sum_{i=1}^{\infty} \mathbb{1}_{A_i} = \sum_{i=1}^{\infty} \mathbb{E}\mathbb{1}_{A_i} \qquad \text{Note}^1$$

$$= \sum_{i=1}^{\infty} \left(\sum_{w \in \Omega} \mathbb{1}_{A_i}(w) p(w) \right)$$

$$= \sum_{i=1}^{\infty} \left(\sum_{w \in A_i} p(w) \right) = \sum_{i=1}^{\infty} p(A_i)$$

$$= \sum_{i=1}^{\infty} (5/6)^{i-1}(1/6) = (1/6)\frac{1}{1 - 5/6} = 1$$

---

[1] $\mathbb{E}\sum_{i=1}^{\infty} X_i = \sum_{i=1}^{\infty} \mathbb{E}X_i$ only if $\sum_{i=1}^{\infty} \mathbb{E}|X_i|$ converges by dominated convergence theorem.

Conditional Independence

Prove or disprove (by counterexample):

(a) $X \perp Y | Z \Rightarrow X \perp Y$

False. Assume that: $P(X = 1) = 0.1$, $P(Y = 1) = 0.5$, $P(Z = 1) = 0.5$, and:

$P(\cdot \mid Z = z)$	$Z = 0$	$Z = 1$
$X = 1$	0.4	0.3
$Y = 1$	0.6	0.2

$$P(X, Y) = \sum_Z P(X, Y, Z) = \sum_Z P(X|Z)P(Y|Z)P(Z)$$

$$P(X, Y) = P(X|Y)P(Y)$$

$$P(X = 1, Y = 1) = 0.4 \times 0.6 \times 0.5 + 0.3 \times 0.2 \times 0.5 = 0.15$$

$$P(X = 1|Y = 1) = P(X = 1, Y = 1)/P(Y = 1) = 0.3 > P(X = 1)$$

Conditional Independence

Prove or disprove (by counterexample):

(b) $(X \perp Y|Z)\&(X \perp Z|Y) \Rightarrow X \perp (Y,Z)$

$$
\begin{aligned}
P(X,Y,Z) &= P(X,Y|Z)P(Z) = P(X|Z)P(Y|Z)P(Z) \\
&= P(X|Z)P(Y,Z) \\
&= P(X,Z|Y)P(Y) = P(X|Y)P(Z|Y)P(Y) \\
&= P(X|Y)P(Y,Z) \\
\Rightarrow & P(X|Z) = P(X|Y)
\end{aligned}
$$

$$
\begin{gathered}
P(X|Z)P(Z)P(Y) = P(X,Z)P(Y) = \\
P(X|Y)P(Y)P(Z) = P(X,Y)P(Z)
\end{gathered}
$$

$$
\underbrace{\sum_z P(X,Z=z)}_{=P(X)} P(Y) = P(X,Y) \underbrace{\sum_z P(Z=z)}_{=1} \Rightarrow X \perp Y
$$

Index

Definition and facts

A vector-valued RV $x \in \mathbb{R}^n$ is said to have a multivariate normal distribution with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in S^n_{++}$ if its pdf is:

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

1. If you know mean and covariance of a Gaussian random variable, you know the whole distribution.
2. Sum of independent gaussians is Gaussian.
3. Marginal of a joint Gaussian is Gaussian.
4. Conditional of a joint Gaussian is Gaussian.

Linear transformations

For $X \sim \mathcal{N}(\mu, \Sigma)$. By factorizing the covariance matrix as $\Sigma = U \Delta U^T = BB^T$, the RV $Z = B^{-1}(X - \mu) \sim \mathcal{N}(0, I)$.

- Proof by change of variables formula:

$$p_Z(z) = p_X(x) \cdot \left| \det \left(\frac{\partial x_i}{\partial z_j^T} \right) \right|$$

- Any gaussian variable can be decomposed into n independent gaussian variables.
- By generating n independent gaussians and applying $BZ + \mu$ any gaussian distribution can be generated.

Diagonal Covariance Case

Consider $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$

$$\begin{aligned} p(x) &= \frac{1}{(2\pi)^{n/2}|\sigma_1^2\sigma_2^2|^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 \right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right) \end{aligned}$$

In general, when Σ is diagonal, then the components of x are indepentent of each other.

Shape of Level sets

A level set of a function $f : \mathrm{R}^n \to \mathrm{R}$ is a set

$$\{x \in \mathrm{R}^n : f(x) = c\},$$

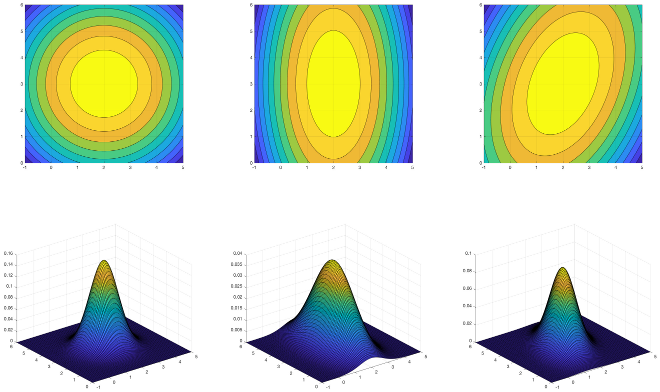
for some $c \in \mathrm{R}$. For 2-D gaussians with diagonal covariance matrix

$$c = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2\right)$$

$$1 = \frac{(x_1 - \mu_1)^2}{2\sigma_1^2 \log\left(\frac{1}{2\pi c\sigma_1\sigma_2}\right)} + \frac{(x_2 - \mu_2)^2}{2\sigma_2^2 \log\left(\frac{1}{2\pi c\sigma_1\sigma_2}\right)}$$

Shape of Level sets

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$



Sum of Independent Gaussians

Assume that $x \sim \mathcal{N}(\mu_x, \Sigma_x)$ and $x \sim \mathcal{N}(\mu_y, \Sigma_y)$ are independent, then $z = x + y$ is also Gaussian (Not proven). Let's calculate it first two moments.

$$\mathbb{E}[z_i] = \mathbb{E}[x_i + y_i] = \mathbb{E}[x_i] + \mathbb{E}[y_i] = \mu_x + \mu_y$$

$$\begin{aligned}
\mathbb{E}[(z_i - \mu_i)(z_j - \mu_j)] &= \mathbb{E}[z_i z_j] - \mathbb{E}[z_i]\mathbb{E}[z_j] \\
&= \mathbb{E}[(x_i + y_i)(x_j + y_j)] - \mathbb{E}[x_i + y_i]\mathbb{E}[x_j + y_j] \\
&= \mathbb{E}[x_i x_j + x_i y_j + x_j y_i + y_i y_j] - \mathbb{E}[x_i + y_i]\mathbb{E}[x_j + y_j] \\
&= \underbrace{\mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j]}_{=\Sigma_{x_{i,j}}} + \underbrace{\mathbb{E}[y_i y_j] - \mathbb{E}[y_i]\mathbb{E}[y_j]}_{=\Sigma_{y_{i,j}}} \\
&\quad + \underbrace{\mathbb{E}[x_i y_j]}_{=\mathbb{E}[x_i]\mathbb{E}[y_j]} - \mathbb{E}[x_i]\mathbb{E}[y_j] + \underbrace{\mathbb{E}[y_i x_j]}_{=\mathbb{E}[y_i]\mathbb{E}[x_j]} - \mathbb{E}[y_i]\mathbb{E}[x_j] \\
&= \Sigma_{x_{i,j}} + \Sigma_{y_{i,j}}
\end{aligned}$$

Marginal of joint Gaussians

$$p(x_A, x_B) = \frac{1}{Z} \exp\left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right)$$

$$V = \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1}, \quad \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} = \begin{bmatrix} \Delta_A \\ \Delta_B \end{bmatrix}$$

$$p(x_A) = \frac{1}{Z} \int_{x_B} \exp\left(-\frac{1}{2} \begin{bmatrix} \Delta_A \\ \Delta_B \end{bmatrix}^T \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} \begin{bmatrix} \Delta_A \\ \Delta_B \end{bmatrix} \right) dx_B, \qquad (\text{Note})^2$$

$$= \frac{1}{Z} \exp\left(-\frac{1}{2} \left[\Delta_A^T (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA}) \Delta_A \right] \right)$$

$$\cdot \int_{x_B} \exp\left(-\frac{1}{2} \left[(\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A)^T V_{BB} (\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A) \right] \right) dx_B$$

$$p(x_A) = \frac{1}{Z_A} \exp\left(-\frac{1}{2} \left[\Delta_A^T (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA}) \Delta_A \right] \right)$$

$$= \frac{1}{Z_A} \exp\left(-\frac{1}{2} \left[\Delta_A^T \Sigma_{AA}^{-1} \Delta_A \right] \right)$$

---

[2] $\frac{1}{2} z^T A z + b^T z + c = \frac{1}{2} (z + A^{-1} b)^T A (z + A^{-1} b) + c - b^T A^{-1} b$

Conditional of joint Gaussians

$$
\begin{aligned}
p(x_B|x_A) =& \frac{p(x_A, x_B)}{p(x_A)} \\
=& \frac{1}{Z'} \exp\left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) \\
=& \frac{1}{Z'} \exp\left(-\frac{1}{2} \left[\Delta_A^T (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA}) \Delta_A \right] \right) \\
& \cdot \exp\left(-\frac{1}{2} \left[(\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A)^T V_{BB} (\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A) \right] \right) \\
=& \frac{1}{Z''} \exp\left(-\frac{1}{2} \left[(\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A)^T V_{BB} (\Delta_B + V_{BB}^{-1} V_{BA} \Delta_A) \right] \right) \\
x_B|x_A \sim& \mathcal{N}(\underbrace{\mu_B - V_{BB}^{-1} V_{BA}(x_A - \mu_A)}_{=\mu_{B|A}}; \underbrace{\Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}}_{=\Sigma_{B|A} = V_{BB}^{-1}})
\end{aligned}
$$