Lecture 3

Probability - Part 3

Luigi Freda

ALCOR Lab
DIAG
University of Rome "La Sapienza"

May 20, 2019

Outline

Outline

Transformation of Random Variables

- if $\mathbf{x} \sim p_{\mathbf{X}}()$ is some random variable, and $\mathbf{y} = f(\mathbf{x})$, what is the distribution of $\mathbf{y}$?
- suppose $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ we can immediately compute

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

  where $\mathbb{E}[x] = \boldsymbol{\mu}$

$$\text{cov}[y] = \mathbb{E}[(\mathbf{A}\mathbf{x} + \mathbf{b} - \mathbb{E}[y])(\mathbf{A}\mathbf{x} + \mathbf{b} - \mathbb{E}[y])^T] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

  where $\text{cov}[x] = \boldsymbol{\Sigma}$
- but what is the PDF $p_{\mathbf{Y}}()$?

Outline

Transformation of Random Variables

Univariate Discrete RVs

- consider a **discrete** RV X with **PMF** $p_X(x)$ and a transformation $y = f(x)$
- one has

$$p_Y(y) = \sum_{x:\ y=f(x)} p_X(x)$$

- suppose X is a discrete RV with $x \in \{1, 2, ..., 10\}$ and $p_X(x) = 1/10$ for each x
- assume the transformation is

$$y = f(x) = \begin{cases} 1 & \text{if } x \text{ is even} \\ 0 & \text{if } x \text{ is odd} \end{cases}$$

  then

$$p_Y(y = 1) = \sum_{x \in \{2,4,6,8,10\}} p_X(x)$$

- N.B.: f is a many-to-one function but in this case we can simply enumerate the favorable events and sum their probabilities (use the first equation above)

Transformation of Random Variables

Univariate Continuous RVs

- consider a **continuous** RV X with **PDF** $p_X(x)$ and **CDF** $F_X(x)$
- assume we are given a transformation $y = f(x)$
- in this case we can compute the CDF of Y

$$F_Y(y) \triangleq P_Y(Y \leq y) = P_Y(f(X) \leq y) = P_X(\{x \in \mathbb{R} : \ f(x) \leq y\})$$

and then compute the PDF $p_Y(y) \triangleq \dfrac{dF_Y}{dy}$ (assuming F is differentiable)

- assume the **transformation** $y = f(x)$ is **invertible**, in particular **strictly increasing**
- in this case we can compute $x = f^{-1}(y)$ and

$$F_Y(y) \triangleq P_Y(f(X) \leq y) = P_X(X \leq f^{-1}(y)) = F_X(f^{-1}(y))$$

i.e. we can compute the PDF of Y from the PDF of X

- taking the derivatives

$$p_Y(y) \triangleq \frac{dF_Y}{dy} = \frac{dF_X(f^{-1}(y))}{dy} = \frac{dF_X}{dx}\bigg|_{x=f^{-1}(y)} \frac{df^{-1}(y)}{dy}$$

- assume the **transformation** $y = f(x)$ is **invertible**, in particular **strictly decreasing**
- in this case

$$F_Y(y) \triangleq P_Y(f(X) \leq y) = P_X(X \geq f^{-1}(y)) = 1 - F_X(f^{-1}(y))$$

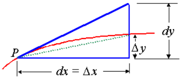
again, we can compute the PDF of Y from the PDF of X

- taking the derivatives

$$p_Y(y) \triangleq \frac{dF_Y}{dy} = -\frac{dF_X(f^{-1}(y))}{dy} = -\frac{dF_X}{dx}\bigg|_{x=f^{-1}(y)} \frac{df^{-1}(y)}{dy}$$

- in both cases, when $y = f(x)$ is **invertible**

$$p_Y(y) = \frac{dF_X}{dx}\bigg|_{x=f^{-1}(y)} \left|\frac{df^{-1}(y)}{dy}\right|$$



more intuitively, one has

$$P_Y(|Y - y| < dy) \approx p_Y(y)|dy| = p_X(x)|dx| \approx P_X(|X - x| < dx)$$

where dy corresponds[1] to dx , hence

$$p_Y(y) = p_X(x)\left|\frac{dx}{dy}\right|$$

[1]note that $\frac{dy}{dx} \lessgtr 0$

Transformation of Random Variables

An example

what happens when the **transformation** $y = f(x)$ is **non-invertible**?

- consider a transformation $y = x^2$
- in this case, let's reason again on the CDF

$$F_Y(y) = P_Y(Y \leq y) = P_X(X^2 \leq y) =$$

$$= P_X(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

- taking the derivatives

$$p_Y(y) = p_X(\sqrt{y})\frac{d}{dy}(\sqrt{y}) - p_X(-\sqrt{y})\frac{d}{dy}(-\sqrt{y}) =$$

$$= \frac{1}{2\sqrt{y}}\Big(p_X(\sqrt{y}) + p_X(-\sqrt{y})\Big)$$

- in general, if f is not invertible, one should start reasoning about the CDF F_Y

homework: ex 2.17 on the book

Outline

- consider two RVs $\mathbf{X}$ and $\mathbf{Y}$ whose values are respectively $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^D$
- assume we have the **PDF** $p_{\mathbf{X}}(\mathbf{x})$ and a transformation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ where $\mathbf{f} : \mathbb{R}^D \rightarrow \mathbb{R}^D$

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_D(\mathbf{x}) \end{bmatrix}$$

- the **Jacobian matrix** of $\mathbf{f}$ is defined as

$$\mathbf{J}_{\mathbf{f}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial(f_1, ..., f_D)}{\partial(x_1, ..., x_D)} \triangleq \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_D}{\partial x_1} & \cdots & \frac{\partial f_D}{\partial x_D} \end{bmatrix}$$

- $|\det(\mathbf{J}_f)|$ measures how much a unit cube changes in volume when we apply $\mathbf{f}$

- if f is an **invertible mapping** and is continuously differentiable, we can define the PDF of the transformed variables by using the Jacobian of the inverse mapping $\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y}) = \mathbf{g}(\mathbf{y})$

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right| = p_{\mathbf{X}}(\mathbf{x}) |\det(\mathbf{J}_g)| \qquad (\text{with } \mathbf{x} = \mathbf{g}(\mathbf{y}))$$

- the volume element $dx_1 dx_2 ... dx_D$ is mapped in the new space to the volume element $|\det(\mathbf{J}_g)| dy_1 dy_2 ... dy_D$ i.e.

$$p_{\mathbf{X}}(\mathbf{x}) dx_1 ... dx_D \ \rightarrow \ p_{\mathbf{X}}(\mathbf{g}(\mathbf{y}))) |\det(\mathbf{J}_g)| dy_1 ... dy_D$$

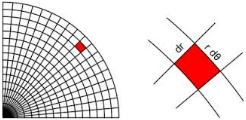
N.B.: the first formula above can be compared to the one used in the scalar case, namely $p_Y(y) = p_X(x) |\frac{dx}{dy}|$

Transformation of Random Variables

Multivariate RVs

a simple example with the polar coordinate transformation

- $\mathbf{x} = (x_1, x_2)^T$ and $\mathbf{y} = (r, \theta)^T$
- $\mathbf{x} = \mathbf{g}(\mathbf{y}) = (r\cos\theta, r\sin\theta)^T$
- one has
$$\mathbf{J_g} = \frac{\partial(g_1, g_2)}{\partial(r, \theta)} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$$
and
$$|\det(\mathbf{J_g})| = |r\cos^2\theta + r\sin\theta^2| = |r|$$
- from $p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(\mathbf{x})|\det(\mathbf{J}_g)|$ it follows
$$p_{r,\theta}(r, \theta) = p_{x_1,x_2}(x_1, x_2)r = p_{x_1,x_2}(r\cos\theta, r\sin\theta)r$$
- in this example the area element $dx_1 dx_2$ is mapped to an area element $rd\theta dr$
- $p_{x_1,x_2}(dx_1, dx_2)dx_1 dx_2 \;\;\rightarrow\;\; p_{x_1,x_2}(r\cos\theta, r\sin\theta)rdrd\theta$



Outline

Central Limit Theorem

central limit theorem

- consider N RVs X_i which are **independent** and **identically distributed** (**iid**)
- i.e., $X_i \sim p(x)$ for $i \in \{1, ..., N\}$ and $p(x_1, ..., x_N) = p(x_1)...p(x_N)$
- let $\mu \triangleq \mathbb{E}[X_i]$ and $\sigma^2 \triangleq \text{var}[X_i]$
- let $\overline{X} \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i$ (note that $\mathbb{E}[\overline{X}] = \mathbb{E}[X_i] = \mu$)
- let $Z_N \triangleq \frac{\overline{X}-\mu}{\sigma/\sqrt{N}}$ then

$$Z_N \xrightarrow{d} \mathcal{N}(0,1) \qquad N \to \infty$$

- a variant of this theorem (due to Lyapunov) states that the sum of independent RVs (NOT identically distributed) with finite means and variances converge to a normal distribution (under certain mild conditions)

Outline

Monte Carlo Approximation

- let $y = f(x)$ be a given RVs transformation
- let $x_1, ..., x_N$ be N samples of the RV X
- **Monte Carlo approximation**: we can approximate the distribution of $Y = f(X)$ by using the empirical distribution of $\{f(x_i)\}_{i=1}^N$

$$p(y) = \sum_{i=1}^{N} w_i \delta_{y_i}(y) \qquad (y_i = f(x_i))$$

- in this way we can approximate

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

- the accuracy of the Monte Carlo estimates increases with the number N of samples

Monte Carlo Approximation

in particular we have

- the mean

$$\mathbb{E}[X] \approx \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- the variance

$$\text{var}[X] \approx \overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

- the median

$$\text{median}(X) = \text{median}\{x_1, ..., x_N\}$$

note that

- $\overline{x} = \arg\min_m \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2$
- $\text{median}\{x_1, ..., x_N\} = \arg\min_m \frac{1}{N}\sum_{i=1}^{N}|x_i - m|$

Monte Carlo Approximation

consider N independent samples with $\mathbb{E}[X_i] = \mathbb{E}[X] = \mu$ and $\text{var}[X_i] = \text{var}[X] = \sigma^2$

- one has

$$\mathbb{E}[\overline{x}] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} x_i] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[x_i] = \mu$$

- more over

$$\text{var}[\overline{x}] = \mathbb{E}[(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu)(\frac{1}{N}\sum_{j=1}^{N} x_j - \mu)] = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[(x_i - \mu)(x_j - \mu)]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}[(x_i - \mu)(x_i - \mu)] = \frac{\sigma^2}{N}$$

where $\mathbb{E}[(x_i - \mu)(x_j - \mu)] = 0$ for $i \neq j$ since X_i and X_j are independent

- we can use $\overline{x}$ as an estimate of μ
- the accuracy of the estimate $\overline{x}$ is $\frac{\sigma^2}{N} \approx \frac{\overline{\sigma}^2}{N}$ which improves as $N \to \infty$

Outline

Entropy

Measure of Uncertainty

- the **entropy** of a discrete RV X is a measure of its **uncertainty**

$$\mathbb{H}[X] \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k)$$

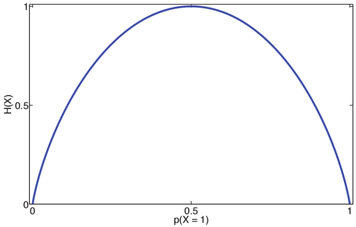
- if you use $\log_2$ the units are **bits**
- if you use $\log$ the units are **nats**
- the discrete distribution with **maximum entropy** is the uniform distribution where $p(x=k) = 1/K$ for any k: in this case $H[X] = \log_2 K$
- the discrete distribution with **minimum entropy** is the delta function $p(x) = \delta_{\overline{k}}(x) = \mathbb{I}(x=\overline{k})$: in this case $H[X] = 0$

N.B.: sometimes the entropy is considered w.r.t. the underlying probability distribution and written as $\mathbb{H}[p]$ instead of $\mathbb{H}[X]$

- let $X \in \{0, 1\}$ has a Bernoulli distribution, i.e. $X \sim \mathrm{Ber}(\theta)$
- $p(X = 1) = \theta$ and $p(X = 0) = 1 - p(X = 1) = 1 - \theta$
- one has

$$\mathbb{H}[X] = -[p(X = 1) \log_2 p(X = 1) + p(X = 0) \log_2 p(X = 0)] =$$
$$= -[\theta \log_2 \theta + (1 - \theta) \log_2 (1 - \theta)]$$

- **binary entropy function**



- the **entropy** of a discrete RV X is a measure of the **information** which is received when we observe a new instance of X

$$\mathbb{H}[X] \triangleq -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k)$$

- in general, the binary representation of a variable with K states needs $\log_2 K$ bits
- consider a RV X which has 8 equally likely states, i.e. $p(X=k) = 1/8$ for any k

$$\mathbb{H}[X] = -8 \cdot \frac{1}{8} \log_2 \frac{1}{8} = \log_2 8 = 3 \text{ bits}$$

  that is exactly the number of bits required to transmit X
- in general, if the states comes with a non-uniform distribution we can use variable length messages using shorter codes for more probable states and longer for less probable states
- **noiseless coding theorem** (Shannon, 1948): the entropy is a **lower bound** on the number of bits needed to transmit the state of a random variable

Entropy

Measure of Dissimilarity

- we can use the entropy concept to measure the **dissimilarity of two probability distributions** $p()$ and $q()$
- **Kullback-Leibler divergence** (KL divergence)

$$\mathbb{KL}[p||q] \triangleq \sum_{i=1}^{K} p_i \log \frac{p_i}{q_i} = -\mathbb{H}[p] + \mathbb{H}[p,q]$$

where $\mathbb{H}[p,q]$ is the **cross entropy**

$$\mathbb{H}[p,q] \triangleq \sum_{i=1}^{K} p_i \log q_i$$

- **theorem**: $\mathbb{KL}[p||q] \geq 0$ with equality **iff** $p = q$
- assume $q_i = 1/K$ (uniform distribution on K states), we have

$$0 \leq \mathbb{KL}[p||q] \triangleq \sum_{i=1}^{K} p_i \log \frac{p_i}{q_i} = -\mathbb{H}[p] + \log K \quad \Longrightarrow \quad \mathbb{H}[p] \leq \log K$$

Entropy
Continuous RV

- the **entropy** of a continuous RV X

$$\mathbb{H}[X] \triangleq -\int_{-\infty}^{+\infty} p(x) \log p(x) dx$$

- this is actually called **differential entropy**

Outline

Mutual Information

- correlation is a very limited measure of dependence
- a more general approach is to determine how similar is a joint distribution $p(X, Y)$ to $p(X)p(Y)$ (recall the definition $X \perp Y$)
- **mutual information** (MI)

$$\mathbb{I}[X; Y] \triangleq \mathbb{KL}[p(X, Y)||p(X)p(Y)] = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

- one has $\mathbb{I}[X; Y] \geq 0$ with equality **iff** $p(X, Y) = p(X)p(Y)$
- **conditional entropy**

$$\mathbb{H}[Y|X] \triangleq -\sum_x \sum_y p(x, y) \log p(y|x) = -\sum_x p(x) \sum_y p(y|x) \log p(y|x)$$

$$= -\sum_x p(x) \mathbb{H}[Y|X = x]$$

this quantifies the amount of information needed to describe the outcome of the RV Y given the value of the RV X

- MI can be interpreted as an **uncertainty reduction** of a variable after observing the other

$$\mathbb{I}[X; Y] = \mathbb{H}[X] - \mathbb{H}[X|Y] = \mathbb{H}[Y] - \mathbb{H}[Y|X]$$

Credits

- Kevin Murphy's book