Lecture 2

Probability theory: basics

Event A : $0 \leq P(A) \leq 1$ is the prob. of A

$$P(\bar{A}) = 1 - P(A)$$

Joint prob. : $P(A \wedge B) = P(A,B)$
"A and B"

If events A & B are indep.,
$$P(A,B) = P(A)\,P(B)$$

$$P(A \vee B) = P(A) + P(B) - P(A,B)$$
"A or B"

Mutually exclusive events: $P(A,B) = 0$,
$$P(A \vee B) = P(A) + P(B).$$

Conditional prob.:
$$P(A,B) = P(B|A)\,P(A), \text{ or}$$
$$P(B|A) = \frac{P(A,B)}{P(A)}.$$

Conditional independence: $P(A,B|C) = P(A|C)\,P(B|C)$
$A \perp B | C$

Discrete random variables:

$x \in \underset{\text{finite or countably infinite}}{X}$

$$p(x) = P(\underset{\substack{\uparrow \\ \text{discrete rv}}}{X} = x), \qquad 0 \le p(x) \le 1, \qquad \sum_{x \in X} p(x) = 1.$$

Continuous random variables:

Cumulative distribution function (cdf):

$$P(x) = P(X \le x) \quad \Rightarrow \quad P(a < X \le b) = P(b) - P(a)$$

Probability density function (pdf):

$$p(x) = \frac{dP(x)}{dx}.$$

Then $P(a < X \le b) = P(b) - P(a) = \int_a^b dx\, p(x)$.

$$P(x < X < x + dx) \approx p(x)\, dx.$$

Marginal distribution:

$$p(x) = \sum_y p(x, y) \quad \text{or}$$

$$p(x) = \int dy\, p(x, y)$$

Product rule: $p(x, y) = p(x)\, p(y|x)$

D variables: $p(x_1, \dots, x_D) = p(x_1)\, p(x_2|x_1) \times p(x_3|x_1, x_2) \times \dots \times p(x_D|x_1, \dots, x_{D-1})$

Moments:

<u>Mean</u> $$E[X] = \sum_x x\,p(x) \quad \text{or} \quad E[X] = \int dx\, x\, p(x)$$

$$E[\underset{\text{const}}{a} X + \underset{\text{const}}{b}] = aE[X] + b$$

$$E\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n E[X_i] \quad \text{, etc.}$$

<u>Variance</u> $$\underset{\underset{\sigma^2}{\equiv}}{V[X]} = E[(X - \overset{\overset{E(X)}{\parallel}}{\mu})^2] = E[X^2] - \mu^2 .$$

$$\mathrm{std}[X] = \sqrt{V(X)} = \sigma .$$

$$V[aX+b] = a^2 V[X] ;$$

for <u>indep.</u> rv we have

$$V\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n V[X_i]$$

<u>Mode</u> $$\vec{x}^* = \underset{\vec{x}}{\operatorname{argmax}}\, p(\vec{x})$$

may not be unique (multimodal distributions)

Conditional expectation:

$$E[X|Y]$$

Note that $E_Y[E_X[X|Y]] =$

$$= E_Y\Big[\int dx\, x\, p(x|y)\Big] =$$

$$\ominus \int dy\, p(y) \left[\int dx\, x\, p(x|y) \right] =$$

$$= \int dx\, x \underbrace{\int dy\, \underbrace{p(x|y)\, p(y)}_{p(x,y)}}_{p(x)} = \underline{\underline{E[X]}}.$$

Bayesian probabilities

Consider $\begin{cases} H = \text{hidden rv (e.g., a model prm)} \\ Y = \text{observed rv} \end{cases}$

Then $p(h,y) = p(h|y)\,p(y) = p(y|h)\,p(h)$, or

$$p(h|y) = \frac{p(y|h)\,p(h)}{p(y)},$$ where

$p(y) = \sum_{h'} p(h', y)$ is the marginal likelihood (aka model evidence)

$$\begin{cases} p(y|h) = \text{likelihood} \\ p(h) = \text{prior} \\ p(h|y) = \text{posterior} \end{cases}$$

Ex. The Monty Hall problem

Doors: 1, 2, 3

a single prize is hidden behind one of the doors.

You select one door (but do not open it). Then the host will open one of the other 2 doors, without the prize behind it. Then you can choose again: stick with your choice or switch to the only other remaining closed door.

Then you get the prize behind your final choice of door.

For example, you initially choose door 1, then the host opens door 3 (with nothing behind it). Should you stick w/door 1 or switch to door 2? Does it make any difference?

Define H_i = hypothesis that prize is behind door $i = 1, 2, 3$

choose $P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}$ for the priors

Conditional likelihoods: [after choosing door 1]

door opened by host Y:

$$\left\{\begin{array}{ll} P(Y=2 \mid H_1) = \frac{1}{2}, & P(Y=3 \mid H_1) = \frac{1}{2} \leftarrow \text{host selects at random} \\ P(Y=2 \mid H_2) = 0, & P(Y=3 \mid H_2) = 1 \\ P(Y=2 \mid H_3) = 1, & P(Y=3 \mid H_3) = 0 \end{array}\right.$$

Similar for the initial choices of door 2 or 3.

Now, consider

$$P(H_i \mid Y=3) = \frac{P(Y=3 \mid H_i)\, P(H_i)}{P(Y=3)}$$

$$P(Y=3) = \sum_{i=1}^{3} P(Y=3 \mid H_i)\, P(H_i) =$$

$$= \frac{1}{3}\left[\frac{1}{2} + 1 + 0\right] = \underline{\underline{\frac{1}{2}}}.$$

So,

$$\begin{cases} P(H_1 \mid Y=3) = \dfrac{1/2 \times 1/3}{1/2} = \dfrac{1}{3}, \\ P(H_2 \mid Y=3) = \dfrac{1 \times 1/3}{1/2} = \dfrac{2}{3}, \\ P(H_3 \mid Y=3) = \dfrac{0 \times 1/3}{1/2} = 0. \quad \leftarrow \text{no car behind door 3} \end{cases}$$

Thus, it's twice as good to switch to door 2 than to stay with door 1 (!)

---

what if $Y=2$?

$$P(Y=2) = \frac{1}{3}\left[\frac{1}{2} + 0 + 1\right] = \frac{1}{2},$$

$$\begin{cases} P(H_1 \mid Y=2) = \dfrac{1/2 \times 1/3}{1/2} = \dfrac{1}{3}, \\ P(H_2 \mid Y=2) = \dfrac{0 \times 1/3}{1/2} = 0, \\ P(H_3 \mid Y=2) = \dfrac{1 \times 1/3}{1/2} = \dfrac{2}{3}. \end{cases}$$

Now it's twice as good to switch to door 3 rather than stay w/door 1.

All other combinations can be obtained by permutation of door indices.

Standard distributions

Bernoulli
$$\begin{cases} p(Y=1) = \theta, \\ p(Y=0) = 1-\theta. \end{cases} \Rightarrow p(y|\theta) = \theta^y (1-\theta)^{1-y}$$

$y \in \{0, 1\}$ $\qquad 0 \leq \theta \leq 1$

Binomial Observe N Bernoulli trials, define $S = \sum_{n=1}^{N} \mathbb{I}(y_n = 1)$ - the total # successes

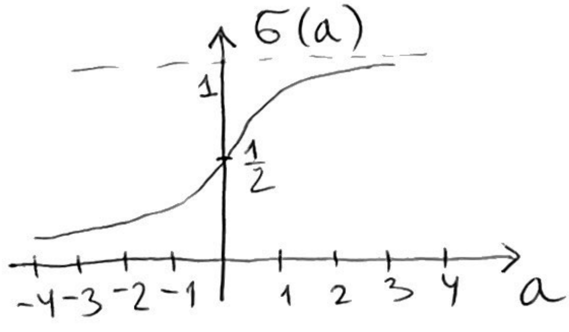
Then $p(S|N,\theta) = \binom{N}{S} \theta^S (1-\theta)^{N-S}$, where

$$\binom{N}{S} = \frac{N!}{(N-S)!\,S!}$$

Sigmoid $C=2$ classifier:

$$p(y|\vec{x},\vec{\theta}) = \underbrace{\mathrm{Ber}}_{\text{Bernoulli}}(y|\underbrace{\sigma(f(\vec{x};\vec{\theta}))}_{\text{sigmoid f'n}})$$

$$\sigma(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{1+e^a}$$



$$\begin{cases} \frac{d}{da}\sigma(a) = \sigma(a)[1-\sigma(a)], \\ 1-\sigma(a) = \sigma(-a) \end{cases}$$

More explicitly,
$$p(y=0) = 1 - \frac{1}{1+e^{-f(\vec{x};\vec{\theta})}} = \frac{1}{1+e^{f(\vec{x};\vec{\theta})}},$$

$$p(y=1) = \frac{1}{1+e^{-f(\vec{x};\vec{\theta})}}$$

$f(\vec{x};\vec{\theta})$ can be any function $\in \mathbb{R}$.

For example, in <u>logistic</u> <u>regression</u>

$$f(\vec{x};\vec{\theta}) = \vec{w}^T \cdot \vec{x} + b .$$

The inverse of $p = \sigma(a)$ is

$$a = \sigma^{-1}(p) = \underbrace{\log\left(\frac{p}{1-p}\right)}_{\text{"log odds"}}$$

Indeed, $\log \dfrac{e^a/(1+e^a)}{1-\dfrac{e^a}{1+e^a}} = \log e^a = \underline{\underline{a}}.$

$$\mathbb{I}(a) = \begin{cases} 1, & a = \text{True} \\ 0, & a = \text{False} \end{cases}$$

<u>Categorical</u> $\quad \text{Cat}(y|\vec{\theta}) = \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)}$

$\vec{\theta} = (\theta_1, \theta_2, \dots, \theta_C)$

$\sum_{c=1}^{C} \theta_c = 1$

$\text{Cat}(y=c|\vec{\theta}) = \theta_c$

One-hot encoding : $y = c \Rightarrow \vec{y} = (0, 0, \dots, 1, 0, \dots)$

@ position c, all other entries $\emptyset$

Then $\quad \text{Cat}(\vec{y}|\vec{\theta}) = \prod_{c=1}^{C} \theta_c^{y_c}$

<u>Multinomial</u>

$$\mathcal{M}(\vec{n}_c \mid N, \vec{\theta}) = \binom{N}{n_1 \ldots n_C} \prod_{c=1}^{C} \theta_c^{N_c},$$

where $N = \sum_{c=1}^{C} n_c$, $\vec{n}_c = (n_1, n_2, \ldots, n_C)$

$$\binom{N}{n_1 \ldots n_C} = \frac{N!}{n_1! n_2! \ldots n_C!}$$

<u>Softmax</u> $C \geq 2$ classifier:

$$p(y = c \mid \vec{x}, \vec{\theta}) = \mathrm{Cat}(y = c \mid \mathrm{softmax}(\vec{f}(\vec{x};\vec{\theta}))) =$$

$$= \mathrm{softmax}(\vec{f}(\vec{x};\vec{\theta}))_c \,.$$

Here, $\underset{(a_1, a_2, \ldots, a_C)}{\mathrm{softmax}(\vec{a})} = \left(\frac{e^{a_1}}{Z}, \ldots, \frac{e^{a_C}}{Z} \right)$,

where

$$Z = \sum_{c'=1}^{C} e^{a_{c'}}.$$

Note that $0 \leq \mathrm{softmax}(\vec{a})_c \leq 1$, and

$$\sum_{c=1}^{C} \mathrm{softmax}(\vec{a})_c = 1 \,.$$

For ex., in <u>multiclass</u> <u>logistic</u> <u>regression</u>

$$\vec{f}(\vec{x};\vec{\theta}) = \overline{W}_{C \times D} \, \vec{x}_D + \vec{b}_C \,, \quad \text{s.t.}$$

$$p(y = c \mid \vec{x};\vec{\theta}) = \mathrm{Cat}(y = c \mid \mathrm{softmax}(\overline{W}\vec{x} + \vec{b})) =$$

$$= \frac{e^{f_c(\vec{x};\vec{\theta})}}{Z} \,.$$