CSEC - 701

Machine Learning

1. Probability Theory basics
2. Linear classification (LDA, SVM)
3. Non-linear classification (QDA, KNN, DT, Boosting)
4. Clustering (Kmeans, EM, Gaussian mixtures, Bayesian Mixtures)
5. Dimensionality Reduction (PCA, ICA, EM, FA)
6. Regression: Linear, polynomial, OLS, EM, logistic, Tobit
7. Feature learning:- NN, Backpropagation, Auto encoder.

Textbook:

A.C. Faul, "A concise introduction to Machine Learning."

	1	2	3	4	5	6	
H	1/12	1/12	1/12	1/12	1/12	1/12	$=\frac{6}{12}=\frac{1}{2}$
T	1/12	1/12	1/12	1/12	1/12	1/12	$=\frac{6}{12}=\frac{1}{2}$
	1/6	1/6	1/6	1/6	1/6	1/6	

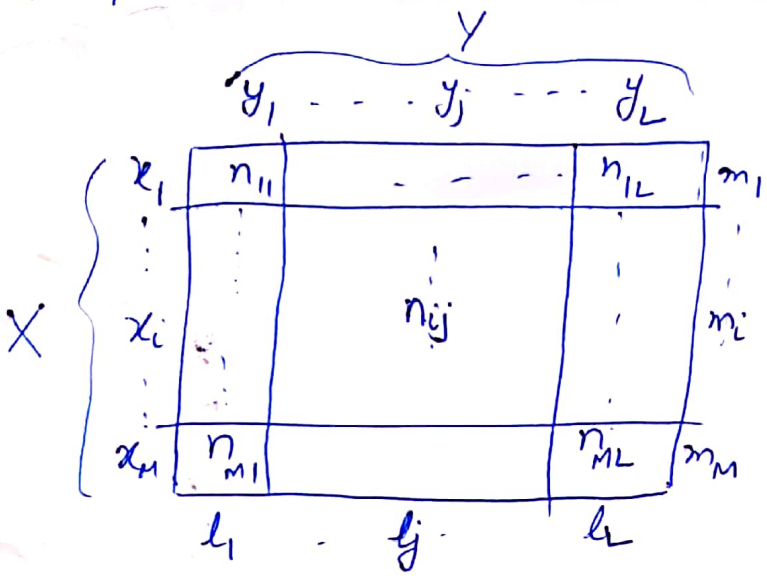
↑ coin

die ↖

Marginal probabilities

Let X can take $x_1, \cdots, x_M \rightarrow P(X=x)$, $x \in \{x_1, \cdots, x_M\}$

" Y " " $y_1, \cdots, y_L \rightarrow P(Y=y)$, $y \in \{y_1, \cdots, y_L\}$



$p(X=x)$ and $p(Y=y)$

$= p(X=x, Y=y)$

$= p(x, y)$

So we use $p(x)$, $p(y)$, $p(x,y)$ for $p(X=x)$, $p(Y=y)$, $p(X=x, Y=y)$

Independance: $p(X,Y) = p(X) \cdot p(Y)$ for all x_i, y_j

$i \in \{1, \cdots, M\}$, $j \in \{1, \cdots, L\}$

Note that $m_i = \sum_{j=1}^{L} n_{ij}$, $l_j = \sum_{i=1}^{M} n_{ij}$

Let Total #Experiments $= N$

$\Rightarrow p(x_i) = \frac{m_i}{N}$, $p(y_j) = \frac{l_j}{N}$, $p(x_i, y_j) = \frac{n_{ij}}{N}$

$$p(x_i) = \sum_{j=1}^{L} p(x_i, y_j)$$

Similarly $p(y_j) = \sum_{i=1}^{M} p(x_i, y_j)$

} Sum rule

Consider only those experiments where $\underbrace{X \text{ is } x_i}_{\text{fix this}}$.

and find prob. $Y = y_j$: $p(Y = y_j | X = x_i)$ or $p(y_j | x_i)$

$$p(y_j | x_i) = \frac{n_{ij}}{m_i}$$

$$p(x_i, y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{m_i} \cdot \frac{m_i}{N} = \underbrace{p(y_j | x_i)\, p(x_i)}_{\text{product rule}}$$

- So Sum rule: $p(X) = \sum_Y p(X, Y)$

  product rule: $p(X, Y) = p(Y | X)\, p(X)$

  joint prob: $p(X, Y) = p(Y, X)$

  $p(Y|X)\, p(X) = p(X|Y)\, p(Y)$ Bayes Rule

- $p(X|Y)$: explains X for several choices of Y. (so function of Y)
  (sometimes called likelihood)

Overall probability of X ($p(X)$) is calculated by using sum and product rules

$$p(X) = \sum_Y p(X|Y)\, p(Y)$$

$p(Y|X)$ is called posterior

(we observed and taken into account)

Ex:

smokers = 582 died = 139

Non-smokers = ~~250~~ 732 died = 230

Mortality rate $= \frac{139}{582} \times 100 = 23.9\%$

rate $= \frac{230}{732} \times 100 = 31.4\%$

	Smoker	Non-Smoker	
died	139	230	369
Alive	443	732	945
	582	732	1314

	18-24		25-34		35-44		45-54		55-64		65-74		75+	
	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS
Dead	2	1	3	5	14	7	27	12	51	40	29	101	13	64
Alive	53	61	121	152	95	114	103	66	64	81	7	28	0	0
	55	62	124	157	109	121	130	78	114	121	36	129	13	64

	18-24	25-34	35-44	45-54	55-64	65-74	75+
Smokers Alive	47%	44%	47%	63%	49%	22%	17%

$\frac{55 \times 100}{55+62}$

$p(\text{died} \mid Y) \qquad Y \in \{S, NS\}$

	18-24	25-34	35-44	45-54	55-64	65-74	75+
Smoker	$\frac{2}{55} \times 100 = 3.6\%$	$\frac{5}{124} \times 100 = 2.4\%$	12.8%	20.8%	44.3%	80.6%	100%
non-Smoker	$\frac{1}{62} \times 100 = 1.6\%$	$\frac{5}{157} \times 100 = 3.2\%$	5.8%	15.4%	33.1%	78.3%	100%

Mortality rates

Gaussian (Normal) Distribution.

$f(x) = \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$ σ^2: variance

$f(x)$ → prob. density function

$\mu = 0, \sigma^2 = 1$ (Standard normal distribution)

Plot distributions for $(\mu,\sigma) \in \left\{(0,1), (0,\frac{1}{2}), (0,2), (2,\frac{1}{\sqrt{2}})\right\}$
Take $x = -10$ to 10

Central limit theorem: $x_1, \cdots, x_n$ sequence of iid random variables with 0 mean and σ^2 variance, then

$$S_n = \frac{x_1 + \cdots + x_n}{\sqrt{n}} \sim \mathcal{N}(x|0,\sigma^2) \text{ for large } n$$

$S_n \to S$ as $n \to \infty$

Entropy.

$$-\int_{-\infty}^{\infty} f(x) \log(f(x))\, dx$$

- Many physical systems strive to maximize the entropy over time.
- For a given μ & σ^2, $f(x) = \mathcal{N}(x|\mu,\sigma^2)$ maximizes the entropy

Properties: $\int_{-\infty}^{\infty} f(x)\, dx = 1$

We want to integrate

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] dx$$

substitute

$$\frac{x-\mu}{\sigma} = y$$

$$dx = \sigma\, dy$$

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{y^2}{2}\right] dy$$

$$I^2 = \left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right]\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}\, dz\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{y^2+z^2}{2}} dy\,dz$$

$y = r\cos\theta$, $z = r\sin\theta$ $\quad dy\,dz = r\,dr\,d\theta$

$$I^2 = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} \exp\left[-\frac{r^2(\cos^2\theta + \sin^2\theta)}{2}\right] r\,dr\,d\theta = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} r e^{-\frac{r^2}{2}} dr\,d\theta$$

$$= \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_0^{\infty} -\left(e^{-r^2/2}(-r)\,dr\right) = -e^{-r^2/2}\Big|_0^{\infty} = +1$$

$\Rightarrow I = 1$

$\pm k\sigma$	1	2	3	4	5
% age data	68.27	95.45	99.73	99.9937	99.99994

Null hypothesis: x does not exist a in the extreme ends.

* The probability of the observed data occuring by chance under the null hypothesis is known as the p-value.

For example $\pm 5\sigma$ case, the probability of data outside this region is $1 - 0.9999994 = 6 \times 10^{-7}$ (for both sides).

~~$p = 0.5 \times 10^{-7}$~~ $p = 3 \times 10^{-7}$ or approximately 1 in 3.5M.

* So, we can set level of significance using σ.
* We can set any predefined threshold value α like $\alpha = 0.01, 0.01, 0.005$ & 0.001 etc.

CDF:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

probability of falling in $(a, b) = F(b) - F(a)$.

$$f(x) = F'(x) = \frac{d}{dx} F(x)$$

for continuous case:-

sum rule $f(x) = \int_{-\infty}^{\infty} f(x, y)\,dy$

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

Bayes' rule $f(y|x) = \frac{f(x|y) f(y)}{f(x)}$

Expectation (1st moment)

$$E[x] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx \qquad x \text{ is continuous.}$$

$$E[x] = \sum_{x} x f(x) \qquad x \text{ is discrete}$$

Variance / 2nd central moment → $Var[x] = E\left[(x - E[x])^2\right]$

$$Var[x] = \int_{-\infty}^{\infty} \left(x^2 - 2x E[x] + E^2[x]\right) f(x)\,dx$$

$$= \underbrace{\int_{-\infty}^{\infty} x^2 f(x)\,dx}_{E[x^2]} - 2E[x] \underbrace{\int_{-\infty}^{\infty} x f(x)\,dx}_{E[x]} + E^2(x) \underbrace{\int_{-\infty}^{\infty} f(x)\,dx}_{1}$$

$$= E[x^2] - E^2[x] \qquad \left(-2E^2[x]\right)$$

Crude moment $\int_{-\infty}^{\infty} f(x)\, x^i\, dx$

ith central moment $\int_{-\infty}^{\infty} f(x)\,(x - E[x])^i\, dx$

Normalized ith central moment (NCM) $= \dfrac{i\text{th central moment}}{\text{var}[x]^i}$

↳ invariant to linear change of the x

i.e., if $y = ax + b$,

$\therefore NCM(y) = NCM(x)$

3rd NCM: skewness — measures asymmetry around the mean of a probability distribution.

symmetric distribution → skewness is zero

More mass on left of mean → skewness < 0

" " right " " → " > 0

4th NCM: kurtosis → how heavily tailed the probability distribution is

large kurtosis → extreme values of random variable are more likely

small kurtosis → outliers are rare