Probability Theory

Frequentist interpretation $p = \frac{f}{N}$ # of times something happen; # of trials

$p(\text{head}) = \frac{1}{2}$

In N tosses we have roughly $N/2$ heads

Bayesian interpretation: p is about our beliefs.

$p(\text{head}) = \frac{1}{2}$ means we believe head and tail are equally likely.

If we observe something to the contrary we will change our opinion.

However the math of probability is still the same in both approaches.

Key ideas needed for Bayesian point of view

Joint probability

$$P(A,B) = P(A \overset{\text{and}}{\cap} B) = \underbrace{P(A|B)}_{\text{Conditional probability}} P(B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Probability distributions; Discrete distribution

Random variable X $\quad x \in \mathcal{X}$

$$p(x) = P(X=x)$$

Joint distribution $P(X=x, Y=y)$

$$P(X=x \mid Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

Bayes' Rule/Theorem!

$$= \frac{P(X=x)\, P(Y=y \mid X=x)}{\sum_{x'} P(X=x')\, P(Y=y \mid X=x')}$$

Examble 1:

$P(y|x)$

Test

Disease x \ y	0	1	
0	0.9	0.1	1
1	0.2	0.8	1
	—	—	

$\times 0.996$

$\times 0.004$

$P(x, y)$

$P(x) = .004$

Test

Disease x \ y	0	1	
0	0.8664	0.0996	0.996
1	.0008	0.0032	0.004
	.8672	.1028	

$P(x|y)$

Test

Disease x \ y	0	1	
0	0.9991	0.9689	—
1	0.0009	0.0311	—
	1	1	

normalize

$$= \frac{0.004 \times 0.8}{0.004 \times 0.8 + .996 \times 0.1}$$

$$= \frac{0.0032}{0.0032 + 0.0996} \approx 0.031$$

Ignoring this prior gives the wrong impressn

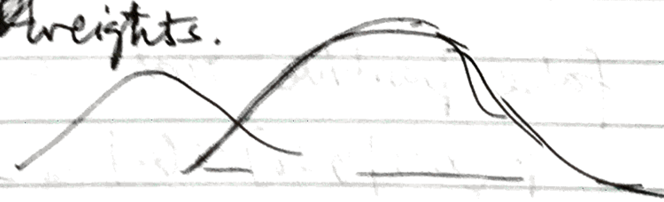
$\Rightarrow$ base rate fallacy.

Example: ?

Generative classifiers

$$p(y=c \mid x) = \frac{P(y=c)\, P(x \mid y=c)}{\sum_{c'} P(y=c')\, P(x \mid y=c')}$$

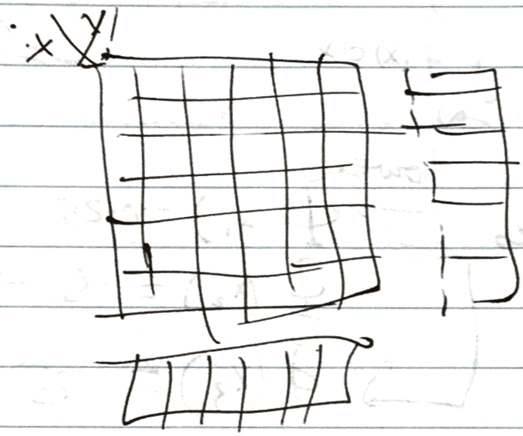
Consider data cloud coming from two normal distribution with different weights.



Independence and conditional independence

$X \perp Y$ means $P(x,y) = P(x,\cdot)P(\cdot,y)$

In that case $P(x|y) = P(x,\cdot)$



$P(x,\cdot)$

marginal distribution

$P(\cdot,y)$

Example:

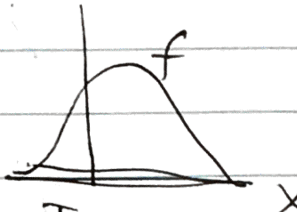
Conditional independence

$X \perp Y | Z$

$P(x,y|z) = P(x,\cdot|z)\, P(\cdot,y|z)$

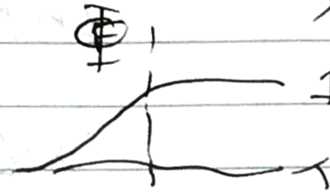
Continuous distribution

$$P(a < X \le b) = \int_a^b f(x)\,dx$$



Cumulative

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



Quartiles

lower $\rightarrow \Phi(x_1) = 0.25$

$\Phi(x_2) = .5 \rightarrow$ Median

upper $\rightarrow \Phi(x_3) = .75$

Mean: $E[X] = \sum_x x\,p(x)$ or $\int_x x\,p(x)\,dx$

Variance: $Var[X] = E[(X - E[X])^2]$

$$= E[X^2] - (E[X])^2$$

$$\sum_x p(x)(x-\mu)^2 = \sum_x p(x)x^2 - 2\sum_x p(x)\,x\,\mu + \mu^2 \sum p(x)$$

$E[X]$?

$$= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2$$
$$= E[X^2] - (E[X])^2$$

$$\text{std}[X] = \sqrt{\text{var}[X]}$$

Some common discrete distribution

Bernoulli distribution

$X \in \{0, 1\}$

$$\text{Ber}(x|\theta) = \theta \quad \text{if } x = 1 \; (\text{success})$$
$$= 1-\theta \quad \text{if } x = 0 \; (\text{failure})$$

like p, $q = 1 - p$

Note that $x^2 = x$

$E[X] = \theta \qquad \text{Var}[X] = ?$

$$\text{Var}[X] = E[X^2] - E[X]^2 = E[X] - E[X]^2$$
$$= \theta - \theta^2 = \theta(1-\theta)$$
$$(= pq)$$

Binomial distribution

$X \sim \text{Bin}(n, \theta) \qquad X \in \{0, \dots, n\}$

successes in n trials.

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$$k = X_1 + \cdots + X_n$$

independent
Bernoulli variables

$E[X] = ?$ $\mathrm{var}[X] = ?$

Related Concept

Prob. Generating Function

$$G(z) = E[z^X] = \sum_{x=0}^{\infty} p(x) z^x$$

Moment Generating Function

$$M_X(t) = E[e^{tX}] = \sum_x p(x) e^{tx}$$

or $\int dx\, f(x) e^{tx}$

Characteristic Function

$$\varphi_X(t) = E[e^{itX}] = \sum_x p(x) e^{itx}$$

or $\int dx\, f(x) e^{itx}$

For Bin(n, θ)

$$G(z) = \sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} z^k = [\theta z + (1-\theta)]^n$$

$$= [1 + \theta(z-1)]^n$$

$$M_X(t) = \{1 + \theta(e^t - 1)\}^n$$

etc..

$$\frac{d^n}{dt^n} M_X(t) = \frac{d^n}{dt^n} \sum_x p(x) e^{tx} = \sum_x p(x) x^n e^{tx}$$

$$\left.\frac{d^n}{dt^n} M_X(t)\right|_{t=0} = \sum_x p(x) x^n = E[X^n]$$

So let us get down to business

$$\frac{d}{dt}\left[1+\theta(e^t-1)\right]^n = n\left[1+\theta(e^t-1)\right]^{n-1} \theta e^t$$

$$\frac{d^2}{dt^2}\left[1+\theta(e^t-1)\right]^n = n\left[1+\theta(e^t-1)\right]^{n-1} \theta e^t$$
$$+ n(n-1)\left[1+\theta(e^t-1)\right]^{n-2} \theta^2 e^{2t}$$

Setting $t=0$ $\quad E[X] = n\theta, \quad E[X^2] = n\theta + n(n-1)\theta^2$

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - E[X]^2$$
$$= n\theta + n(n-1)\theta^2 - n^2\theta^2$$
$$= n(\theta - \theta^2) = n\theta(1-\theta)$$

Known results.

Multinomial distribution

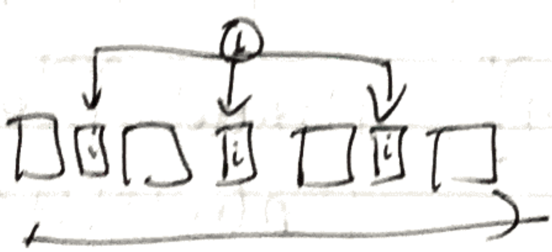
Think of k-sided die with $\theta_1, \cdots, \theta_K$ as the probilities of the different side falling

$$\binom{n}{x_1 \cdots x_k} = \frac{n!}{x_1! \cdots x_k!}$$

$$\sum_{i=1}^{k} x_i = n$$

$$\mathrm{Mu}(x \mid n, \theta) = \binom{n}{x_1 \cdots x_k} \prod_{j=1}^{k} \theta_j^{k}$$

vector



n places

$$\frac{n!}{x_1!\, x_2! \cdots x_i! \cdots x_k!}$$

~~Example 2 DNA~~

$n=1$

$$M(x|1,\theta) = \prod_{j=1}^{k} \theta_j^{I(x_j=1)}$$

$x=(0,0,1,0,0,0)$, for example

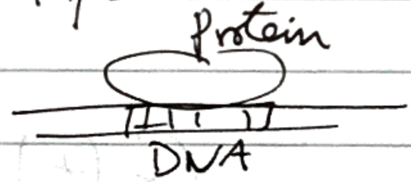
1 hot encoding

Notn: $Cat(x|\theta) \triangleq Mu(x|1,\theta)$

Categorical distribution

Example: DNA Seq. motifs

A, C, G, T



At different locations we have different multinomial distribution

A T T C A
A T C T A
T T T T A
A T A C A

$\downarrow$

$(\theta_{3A}, \theta_{3C}, \theta_{3G}, \theta_{3T})$

$(.25, .25, 0, .5)$

A motif is specified by a matrix with entries θ_{ib}

i: location

b: bases

Poisson distrn

Fun stuff of taking limits

For $\text{Bin}(n, \theta)$ $\quad G(z) = [1 + \theta(z-1)]^n$

If we take the limit $n \to \infty$
with $\theta_n = \frac{\lambda}{n}$

$$G(z) = \left[1 + \frac{\lambda}{n}(z-1)\right]^n$$

$$\to e^{\lambda(z-1)} = G_\lambda(z)$$

$$G_\lambda(z) = \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{n!} z^x$$

This is poisson distrn

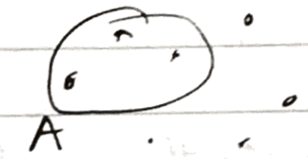
$$\text{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

$$M_x(t) = e^{\lambda(e^t - 1)}$$

$$E[X] = \lambda \qquad \text{Var}[X] = \lambda$$

Empirical distr$^{\underline{n}}$

Data $\mathcal{D} = \{x_1, \dots, x_N\}$



$P_{emp}(A)$

$$= \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(A) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x_i \in A)$$

If we assign weights $w_i \geq 0$ with $\sum w_i = 1$

$$\sum_{i=1}^{N} w_i \, \delta_{x_i}(A)$$

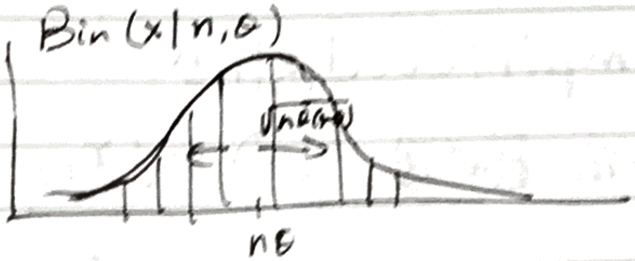
Now it is time to move to continuous distributions.

For that, consider this limit of Binomial distribution

$$z = \frac{x - n\theta}{\sqrt{n\theta(1-\theta)}}$$

What is the distribution of z, when n is large

Note that $E[Z] = 0$ $\quad Var[Z] = 1$



Now note that the characteristic function has the following properties

$$\varphi_{X+a}(t) = E[e^{it(X+a)}] = e^{ita} E[e^{itX}]$$

$$= e^{ita}\varphi_X(t)$$

Also $$\varphi_{cX}(t) = E[e^{itcX}]$$

$$= \varphi_X(ct)$$

If $X \sim Bin(n,\theta)$

$$\varphi_X(t) = E[e^{itX}]$$

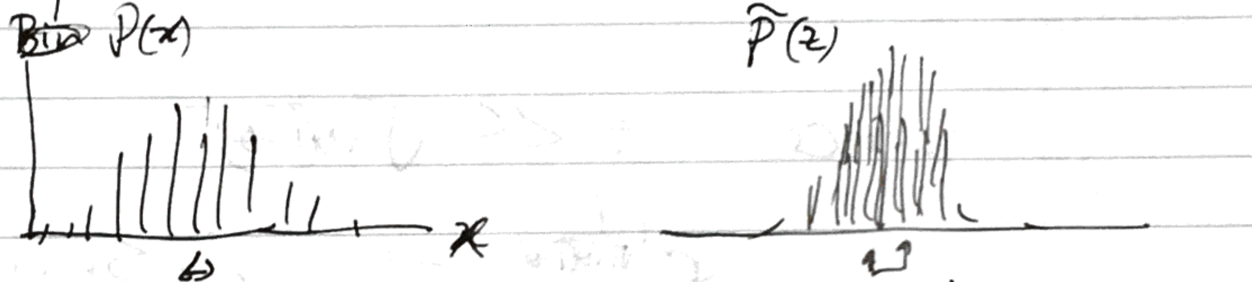
$$= \left(1+\theta(e^{it}-1)\right)^n$$

$$Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

$$\varphi_Z(t) = \varphi_{X-n\theta}\left(\frac{t}{\sqrt{n\theta(1-\theta)}}\right)$$

$$= \varphi_X\left(\frac{t}{\sqrt{n\theta(1-\theta)}}\right) e^{-\frac{in\theta t}{\sqrt{n\theta(1-\theta)}}}$$

So far, this is exact.



Spacing 1 between spikes

spacing $\frac{1}{\sqrt{n\theta(1-\theta)}} \to 0$ as $n \to \infty$

Approximating a continuous distribution as $n \to \infty$.

$\varphi_Z(t)$ is essentially the Fourier transformation of $P(Z)$. The variable t are like momenta.

To produce all this is small scale structure $1/\sqrt{n}$ we need $t \sim \sqrt{n}$. What if we consider $t \sim 1 << \sqrt{n}$.

$$\varphi_z(t) = \left[1 + \theta\left(e^{\frac{it}{\sqrt{n\theta(1-\theta)}}} - 1\right)\right]^n e^{-\frac{in\theta t}{\sqrt{n\theta(1-\theta)}}}$$

$$= e^{n \ln\left[1+\theta\left(e^{\frac{it}{\sqrt{n\theta(1-\theta)}}}-1\right)\right] - \frac{in\theta t}{\sqrt{n\theta(1-\theta)}}}$$

Since $t \ll \sqrt{n\theta(1-\theta)}$

$e^{\frac{it}{\sqrt{n\theta(1-\theta)}}} - 1$ ~ small

$$n \ln\left[1+\theta\left(e^{\frac{it}{\sqrt{n\theta(1-\theta)}}}-1\right)\right] - \frac{in\theta t}{\sqrt{n\theta(1-\theta)}}$$

$$= n\theta\left(e^{\frac{it}{\sqrt{}}} - 1\right) - \frac{n\theta^2}{2}\left(e^{\frac{it}{\sqrt{}}} - 1\right)^2 + n\theta^3(\;)^3 - \frac{in\theta t}{\sqrt{}}$$

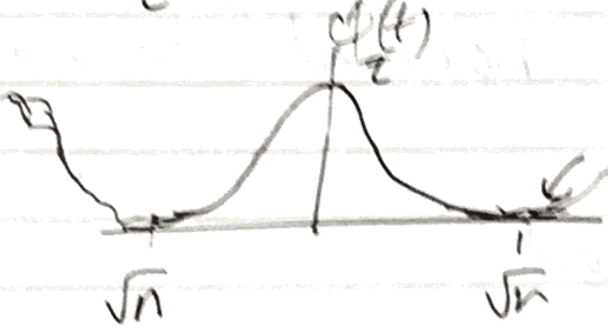
$$= n\theta\left\{\frac{it}{\sqrt{}} + \frac{1}{2}\left(\frac{it}{\sqrt{}}\right)^2 + \dots\right\} - \frac{n\theta^2}{2}\left(\frac{it}{\sqrt{}}\right)^2 - \frac{in\theta t}{\sqrt{}}$$

$$= -\frac{n\theta(1-\theta)}{2}\frac{t^2}{n\theta(1-\theta)} + O\left(\frac{nt^3}{(n\theta(1-\theta))^{3/2}}\right)$$

$$= -\frac{t^2}{2} + O\left(\frac{t^3}{n^{1/2}}\right)$$

So for $|t| \ll n^{1/2}$

$$\varphi_Z(t) \approx e^{-t^2/2}$$



here important for producing discrete spikes

If we only keep $e^{-t^2/2}$, the inverse Fourier transform is going to be smoothe

$$\int_{-\infty}^{\infty} e^{-t^2/2} \frac{e^{-itz}}{2\pi}\, dt$$

$$= \underbrace{\int_{-\infty}^{\infty} e^{-(t+iz)^2/2}\, dt}_{\sqrt{2\pi}} \; \frac{e^{-z^2/2}}{2\pi}$$

$$= \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

That is the probability density function of the standard normal variate

In general

Normal distribution:

Pdf

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E[x] = \mu, \quad Var[x] = \sigma^2$$