ML an Opt Lecture 3

- Joint r.v.s (pdfs, pmfs)
- Estimating pmfs from data
- Marginals
- Conditional distributions
- Independence
- Conditional Independence
- Naive Bayes Classification

Joint random variables (equiv random vector)

$x_1, \dots, x_d, y$ – are joint random variables if they follow a pdf or pmf

$$p(x_1, \dots, x_n, y)$$

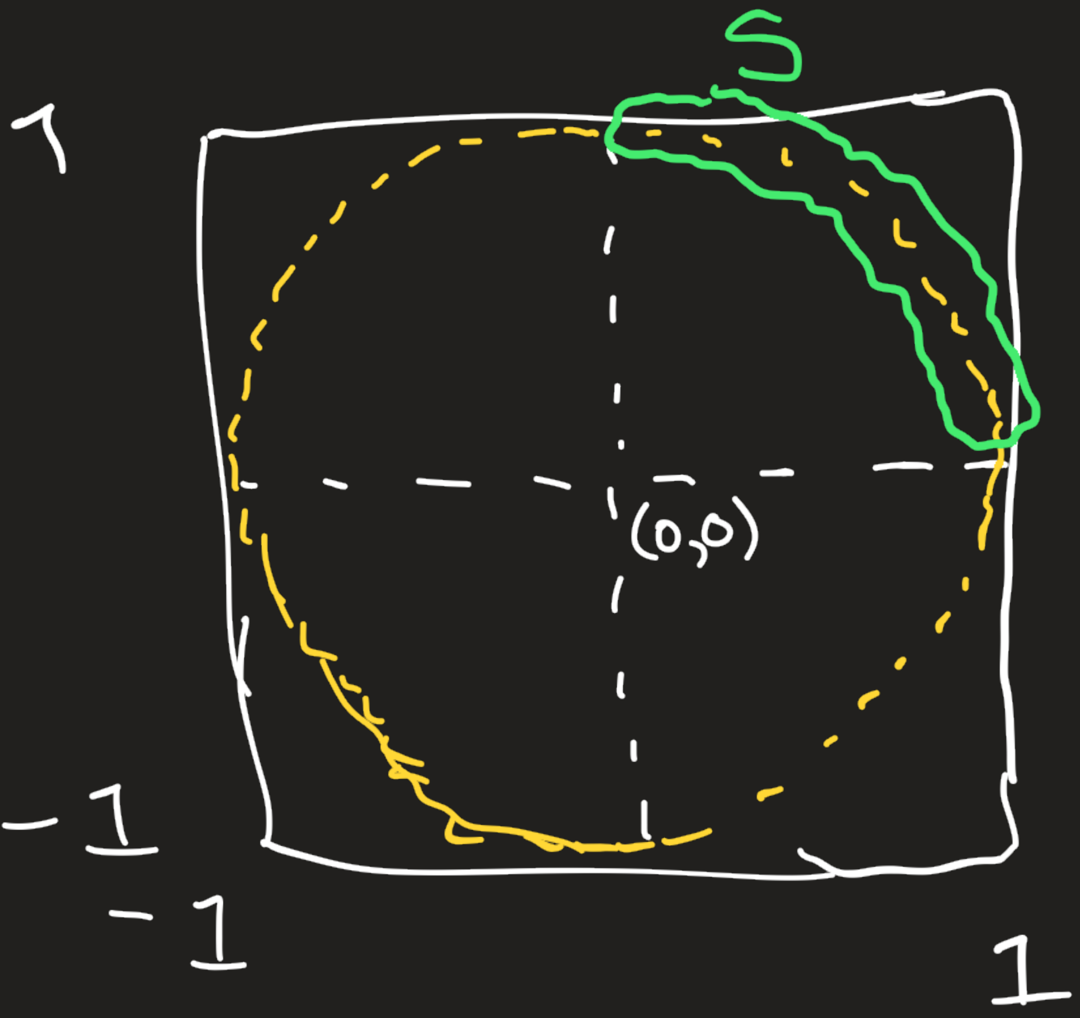
i.e. the probability that $\underset{\overset{\parallel}{z}}{(x_1, \dots, x_d, y)} \in S \subseteq \mathbb{R}^{d+1}$

is given by

$$\mathbb{P}((x_1, \dots, x_d, y) \in S) = \sum_{\omega \in S} p(x_1 = \omega_1, \dots, x_d = \omega_d, y = \omega_{d+1})$$

$$\mathbb{P}(z \in S) = \int_S p(z = \omega)\, d\omega$$

Ex: joint r.v.



$$p(x,y) = \begin{cases} \frac{1}{2\pi} & \text{if } x^2+y^2=1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{P}(S) = \int_S p(\omega)\, d\omega$$

$$= \int_S \frac{1}{2\pi}\, d\omega = \frac{1}{2\pi}\int_S d\omega$$

$$= \frac{1}{2\pi} \cdot \frac{2\pi}{4} = \frac{1}{4}$$

Marginals

If z is a random vector in $\mathbb{R}^d$ governed by pmf p_z, then the marginal pmf/pdf for its ith coordinate z_i is given by

$$p_{z_i}(a) = \sum_{\substack{a_1,\dots,a_{i-1},\\ a_{i+1},\dots,d_d}} p(z_1 = a_1, \dots, z_{i-1} = a_{i-1}, \underline{z_i = a}, z_{i+1} = a_{i+1}, \dots, z_d = a_d)$$

Facts: p_{z_i} is a valid pmf, in that

1) $p_{z_i} \geq 0$ everywhere

2) $\sum_{a_i} p_{z_i}(a_i) = 1$

Exs of marginals

$X \backslash Y$	0	1	2	3	p_X
0	$\frac{1}{16}$	$1/16$	$1/16$	$1/16$	$1/4$
1	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
2	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
p_Y ↑	$1/4$	$1/4$	$1/4$	$1/4$	

$$p_{X,Y}(\omega_1, \omega_2) = \frac{1}{16}$$

$$\Downarrow$$

$$p_X(\omega_1) = \frac{1}{4}$$

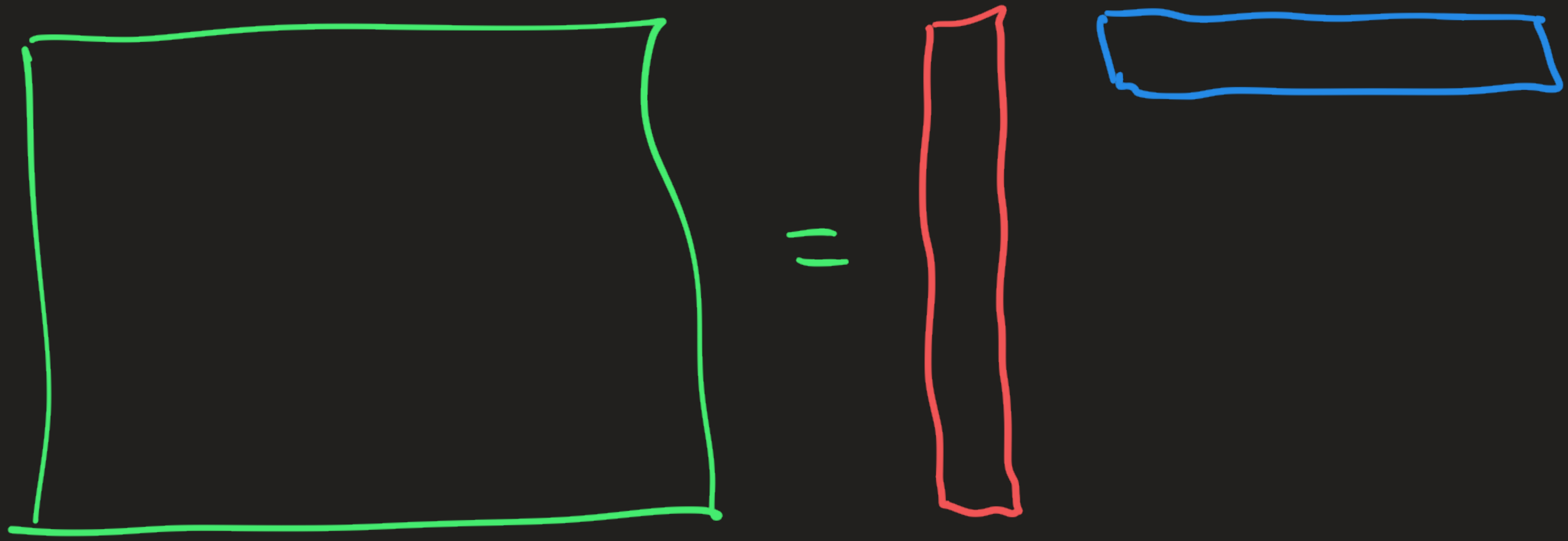
$$p_Y(\omega_2) = \frac{1}{4}$$

We see that $p_{X,Y}(\omega_1, \omega_2) = p_X(\omega_1)\, p_Y(\omega_2)$

X \ Y	0	1	2	3	p_X
0	$\frac{1}{8}$	$7/160$	$7/160$	$1/32$	$39/160$
1	$7/160$	$\frac{1}{8}$	$7/160$	$7/160$	$41/160$
2	$7/160$	$7/160$	$\frac{1}{8}$	$7/160$	$41/160$
3	$\frac{1}{32}$	$7/160$	$7/160$	$\frac{1}{8}$	$39/160$
p_Y	$\frac{39}{160}$	$\frac{41}{160}$	$\frac{41}{160}$	$\frac{39}{160}$	

$$p_{X,Y}(\omega_1,\omega_2) = p_X(\omega_1)\,p_Y(\omega_2) \; ?$$

NO



Conditional distributions

If $z \in \mathbb{R}^d$ is a joint r.v. and I partition it into two sets of joint r.v.s, $z = (z_1, z_2)$

$$z_1 \in \mathbb{R}^{d_1}, \quad z_2 \in \mathbb{R}^{d_2}$$

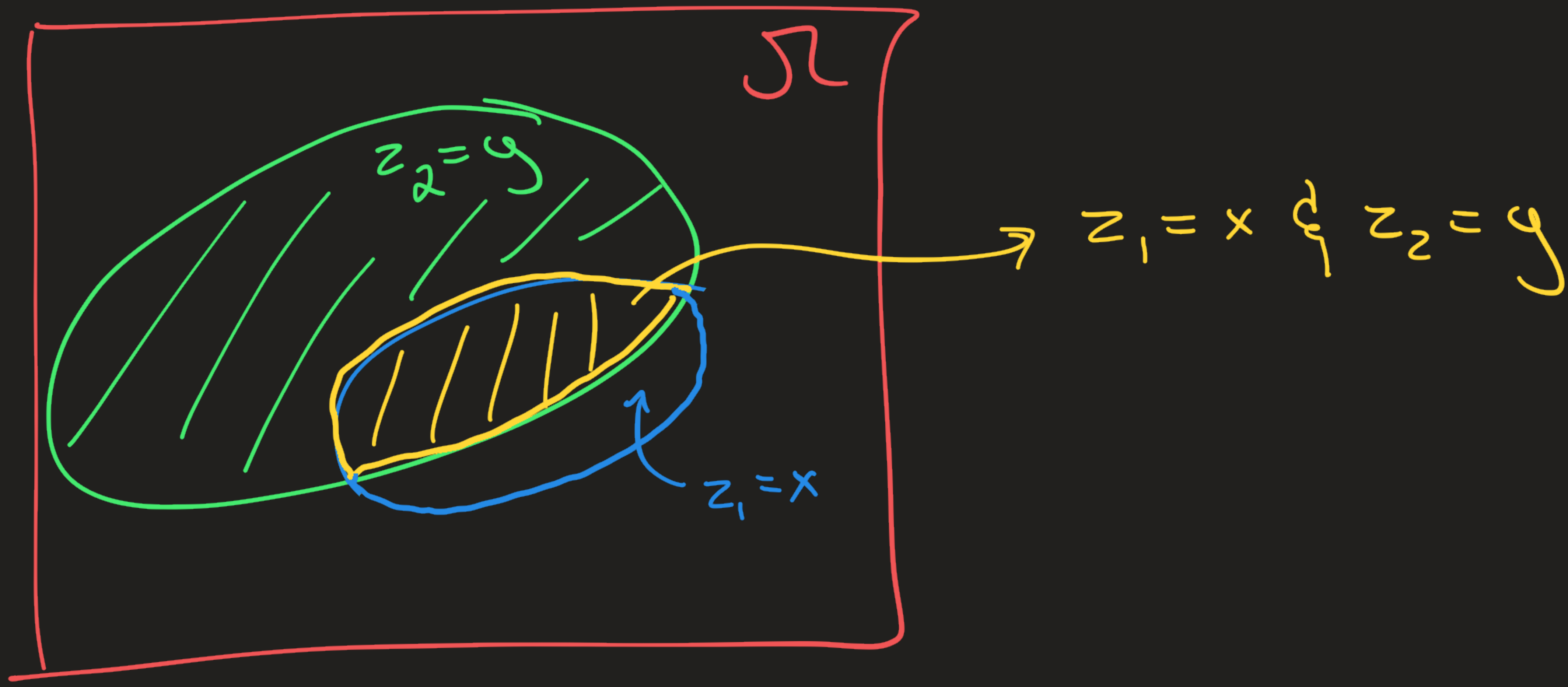
with $d_1 + d_2 = d$

conditional pmf / pdf

then $p(z_1 | z_2)$ captures how knowledge of z_2 modifies our expectation of how z_1 behaves

Def:

$$p(z_1 = x | z_2 = y) = \frac{p_z(x, y)}{p_{z_2}(y)}$$



$$p(z_1 = x \mid z_2 = y) = \frac{p_2(x, y)}{p_{z_2}(y)}$$

$$p(z_1 = x)$$

$z_1 \perp\!\!\!\perp z_2$

$$p_z(x, y) = p_{z_1}(x)\, p_{z_2}(y) \qquad \forall x, y$$

Independence

Two r.v.s x & y are independent if and only if

$$p_{x,y}(\omega_1, \omega_2) = p_x(\omega_1)\, p_y(\omega_2)$$

equivalently

$$p_{x|y}(\omega_1 | \omega_2) = \frac{p_{x,y}(\omega_1, \omega_2)}{p_y(\omega_2)}$$

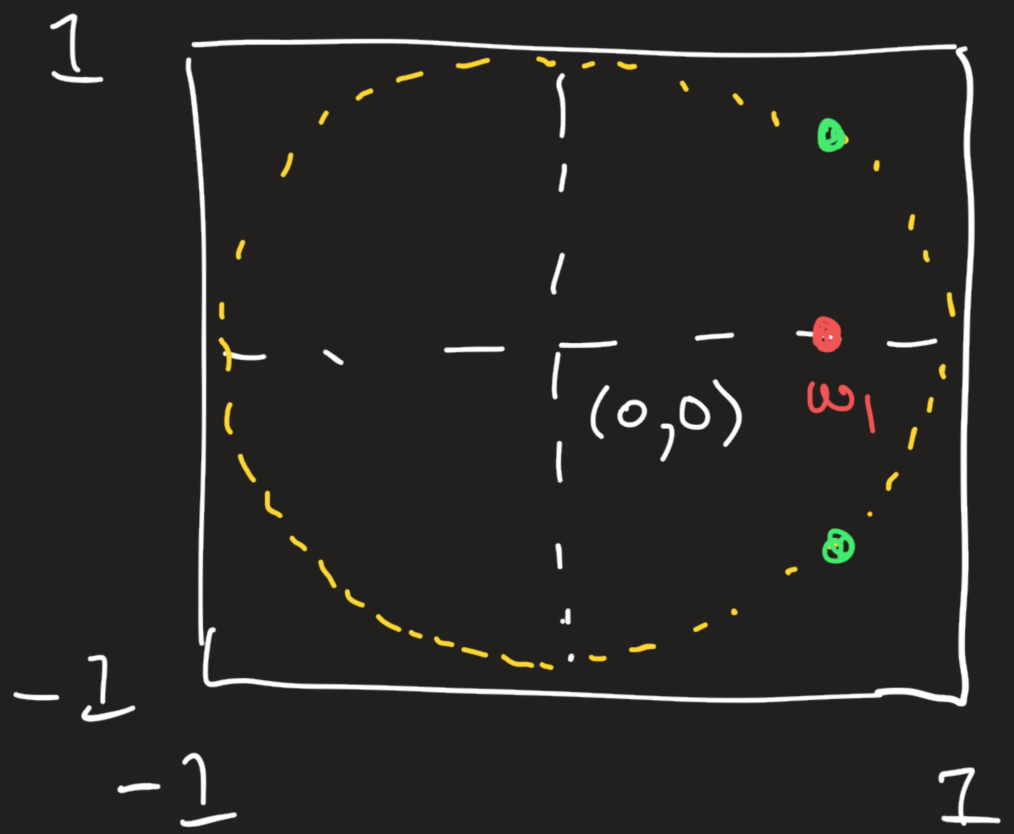
$$= \frac{p_x(\omega_1)\, p_y(\omega_2)}{p_y(\omega_2)} = p_x(\omega_1)$$

that is

$$X \perp\!\!\!\perp Y \iff p_{x,y}(\omega_1, \omega_2) = p_x(\omega_1)\, p_y(\omega_2)$$

$$\iff p_{x|y}(\omega_1 | \omega_2) = p_x(\omega_1)$$

$$\iff p_{y|x}(\omega_2 | \omega_1) = p_y(\omega_2)$$



$$p_{x,y}(\omega_1,\omega_2) = \begin{cases} \dfrac{1}{2\pi} & \text{if } \omega_1^2+\omega_2^2=1 \\ 0 & \text{otherwise} \end{cases}$$

see the end notes for this calculation if you like, otherwise take it by fiat

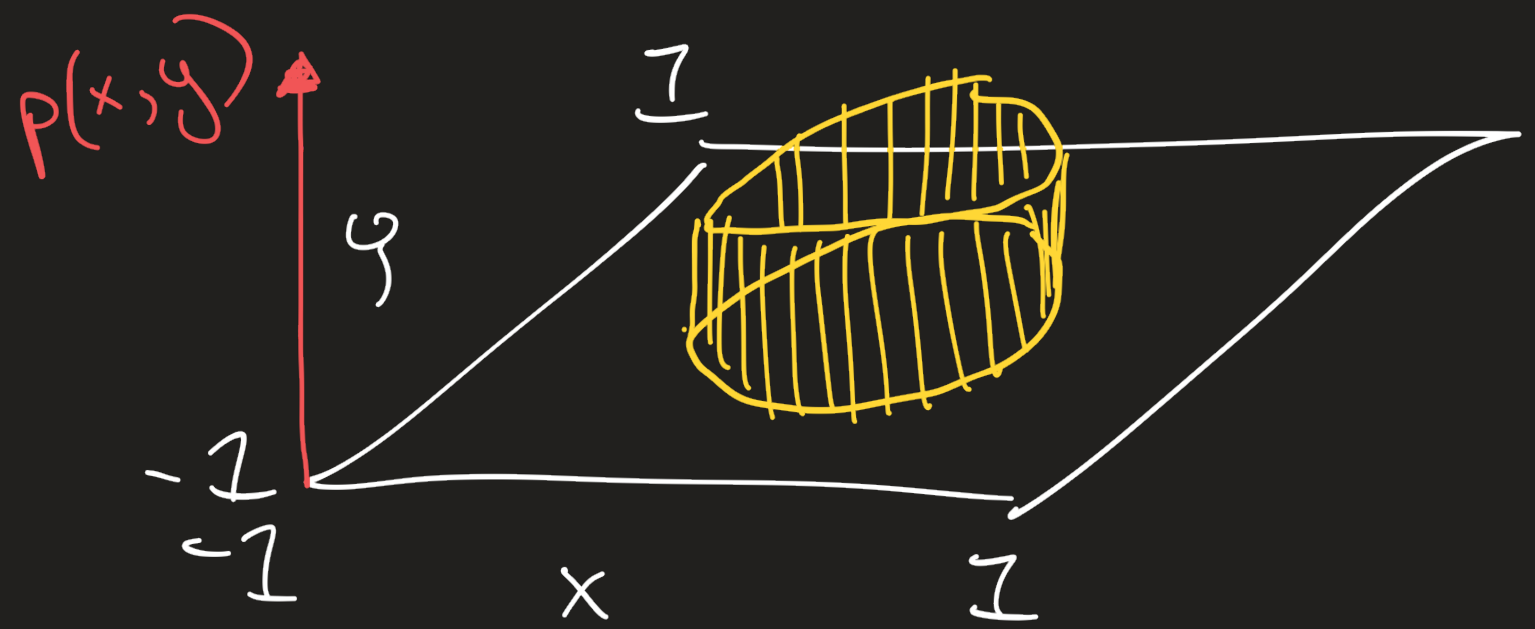
Check if $X \perp\!\!\!\perp Y$:

$$p_x(\omega_1) = \frac{2}{\pi}\sqrt{\frac{1}{1-\omega_1^2}} \quad \text{so similarly} \quad p_y(\omega_2) = \frac{2}{\pi}\sqrt{\frac{1}{1-\omega_2^2}}$$

Clearly then

$$p_{x,y}(\omega_1,\omega_2) \neq p_x(\omega_1)\,p_y(\omega_2)$$

so <u>X & Y are dependent.</u>


p(x, y)
y
1
-1
-1
x
1

Conditional Independence

Maybe X & Y are not independent, but that dependence is mediated by another r.v. Z

Ex:

$$\underset{\text{"intelligence"}}{X} \longrightarrow \underset{\text{"GPA"}}{Z} \longrightarrow \underset{\text{"Salary"}}{Y}$$

Def: $X \perp\!\!\!\perp Y \mid Z$ if

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z)$$

Classification

Goal: given independent r.v.s $x \in \mathbb{R}^d$

predict a class label $y \in \{\pm 1\}$

Ex: spam classification of emails

y = spam/not spam

x = features computed from the email

= (domain of sender, length of emails, entropy, counts of certain n-grams)

Model $p(y|x)$ and use to predict the label

$$\hat{y} = \underset{y \in \{\pm 1\}}{\operatorname{argmax}} \; p(y|x)$$

Problem: $p(y|x)$ can be arbitrarily complicated

Solution: assume a form to $p(y|x)$ to simplify the optimization (this assumption could give good performance in practice, or not)

Naive Bayes assumption

Assume class-conditional independence of the features:

$$p(x|y) = \prod_{i=1}^{d} p(x_i|y)$$

From the NB assumption,

$$p(y=c|x) = \frac{p(x, y=c)}{p(x)} = \frac{p(x|y=c)\,p(y=c)}{p(x)}$$

$$= \frac{p(y=c) \prod_{i=1}^{d} p(x_i|y=c)}{\sum_{l \in \{\pm 1\}} p(x, y=l)}$$

$$= \frac{p(y=c) \prod_{i=1}^{d} p(x_i|y=c)}{p(y=-1)\,p(x|y=-1) + p(y=1)\,p(x|y=1)}$$

so

$$p(y=c|x) = \frac{p(y=c)\prod_{i=1}^{d} p(x_i|y=c)}{p(y=-1)\prod_{i=1}^{d} p(x_i|y=-1) + p(y=1)\prod_{i=1}^{d} p(x_i|y=1)}$$

To maximize $p(y=c|x)$ w.r.t. c, I need to know:

- $p(y=c)$ — prior probabilities of spam/not-spam
- $p(x_i|y=c)$ for $i=1,\dots,d$ — class-conditional probability of the ith observed feature

Can estimate these from the frequencies of the features in data

Endnote: Computing marginal when (x,y) unif on the circle

We said $p_{x,y}(\omega_1,\omega_2) = \begin{cases} \frac{1}{2\pi} & \omega_1^2+\omega_2^2=1 \\ 0 & \text{otherwise} \end{cases}$.

This is not quite mathematically correct, because we can see that

$$\int_A p_{x,y}(\omega_1,\omega_2) = 0 \text{ for any set } A \Rightarrow \mathbb{P}(\Omega)=0$$

which is a contradiction

b/c $p_{x,y}$ is non-zero only on "a set of measure zero".

Really what we mean is

$$(*) \qquad p_{x,y}(\omega_1,\omega_2) = \frac{1}{2\pi}\,\delta(\omega_1^2+\omega_2^2-1)$$

where δ is a special "function" (technically, a tempered distribution) that satisfies

$$\int_{\mathbb{R}} \delta(x) = 1 \quad \text{and} \quad \delta = \begin{cases} 1 & \text{if } x=0 \\ 0 & \text{elsewhere} \end{cases}$$

here "equality" holds in a nonstandard way

(*) is the main necessary technical point.

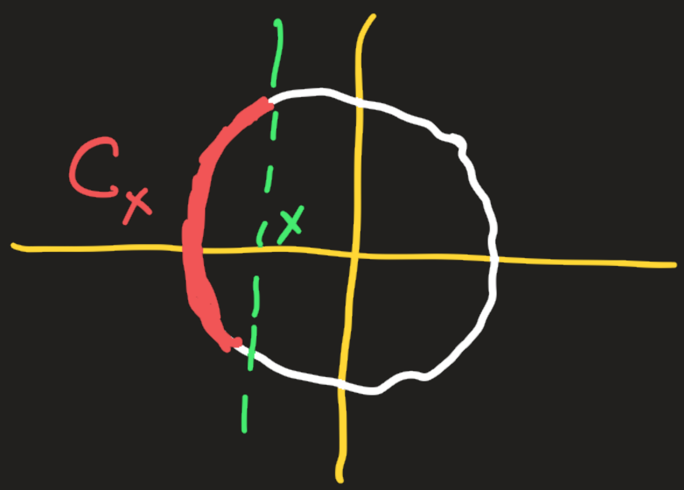
Now we'll use the fact that the pdf of a distribution can be obtained from the derivative of the cdf, cumulative distribution function,

$$F_X(x) \equiv \mathbb{P}(X \le x) = \int_{-\infty}^{x} p_X(t)\, dt$$

This follows from the fundamental theorem of calculus:

$$(**) \quad \frac{d}{dx} F_X(x) = \frac{d}{dx} \mathbb{P}(X \le x) = \frac{d}{dx} \int_{-\infty}^{x} p_X(t)\, dt = p_X(x).$$

In our case, we see that



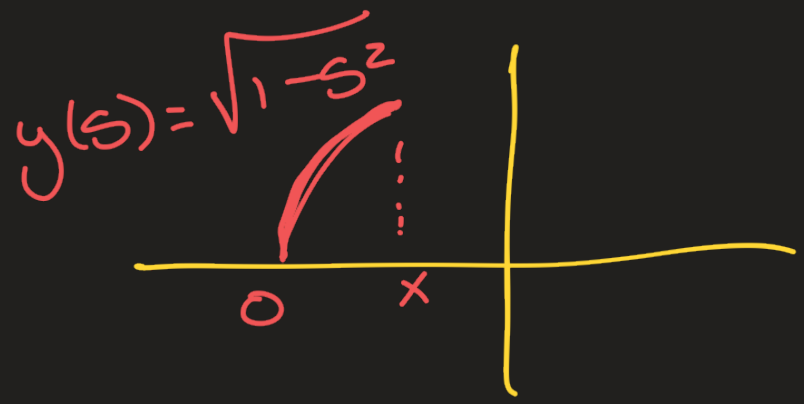
$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(C_x)$$

where $C_x = \{(\omega_1, \omega_2) : \omega_1 \le x \text{ and } \omega_1^2 + \omega_2^2 = 1\}$

and $\mathbb{P}(C_x)$ is the arc-length of C_x divided by 2π.

To compute this note that the half of C_x in the upper half-plane is the graph of the function

$y(s) = \sqrt{1-s^2}$ from $s=-1$ to $s=x$



so from our formulas for arc-lengths of graphs of functions,

$$\mathbb{P}(C_x) = \frac{2}{2\pi}\int_{-1}^{x} \sqrt{1+\left(\frac{dy}{ds}\right)^2}\, ds = \frac{1}{\pi}\int_{-1}^{x}\sqrt{\frac{1}{1-s^2}}\, ds$$

so $\quad p_X(x) = \dfrac{1}{\pi}\sqrt{\dfrac{1}{1-x^2}}$