ML on 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)

$X_1, \ldots, X_d, Y$ are joint random variables if they follow a pdf or pmf

$$p(x_1, \ldots, x_n, y) \equiv$$

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

is given by

$$\mathbb{P}( (x_1, \ldots, x_d, y) \in S ) = \sum_{\omega \in S} p(x_1 = x_1, \ldots, x_d = x_d, y)$$

$$\mathbb{P}( z \in S ) = \int_S p(z = \omega) \, d\omega$$
\[ p(x, y) = \begin{cases} \frac{1}{2\pi} & \text{if } x^2 + y^2 = 1 \\ 0 & \text{otherwise} \end{cases} \]

\[ P(S) = \int_S p(\omega) \, d\omega \]

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

\[ = \frac{1}{2\pi} \cdot 2\pi \cdot \frac{4}{4} = \frac{1}{4} \]
Marginals

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

$$P_{z_i}(a) = \sum_{a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_d} p(z_1 = a_1, \ldots, z_{i-1} = a_{i-1}, z_i = a, z_{i+1} = a_{i+1}, \ldots, z_d = a_d)$$

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

\[
p_{x,y}(\omega_1, \omega_2) = \frac{1}{16}
\]

\[
p_x(\omega_1) = \frac{1}{4}
\]

\[
p_y(\omega_2) = \frac{1}{4}
\]

We see that

\[
p_{x,y}(\omega_1, \omega_2) = \frac{p_x(\omega_1)}{p_y(\omega_2)}
\]
\[
P(x) = \frac{39}{160}
\]
\[
P(y) = \frac{41}{160}
\]
\[
P(x, y) = \frac{39}{160}
\]
\[
P(x) \cdot P(y) = \frac{39}{160} \cdot \frac{41}{160} = \frac{31}{160}
\]

\[
p(x, y) \neq P(x) \cdot P(y)
\]

\[
\text{NO}
\]
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., \( z = (z_1, z_2) \) \( \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \) with \( d_1 + d_2 = d \)

then

the conditional pmf / pdf

\[
p(z_1 | z_2) \text{ captures how knowledge of } z_2 \text{ modifies our expectation of how } z_1 \text{ behaves}
\]

Def:

\[
p(z_1 = x | z_2 = y) = \frac{p_{z_2}(x, y)}{p_{z_2}(y)}
\]
\[ p(z_1 = x \mid z_2 = y) = \frac{p_{Z_1}(x, y)}{p_{Z_2}(y)} \]

\[ p(z_1 = x) \]
$Z_1, \ldots, Z_2$

$P_{Z_2}(x, y) = P_{Z_1}(x) P_{Z_2}(y)$
Independence

Two r.v.s \( X \) and \( Y \) 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

\[
X \perp \perp Y \iff P_{X,Y}(\omega_1, \omega_2) = P_X(\omega_1) P_Y(\omega_2)
\]

\[
\iff P_{X|Y}(\omega_1 | \omega_2) = P_X(\omega_1)
\]

\[
\iff P_{Y|X}(\omega_2 | \omega_1) = P_Y(\omega_2)
\]
Check if $X \perp Y$:

$$p_X(\omega_1) = \frac{2}{\pi} \sqrt{\frac{1}{1-\omega_1^2}}$$

so similarly:

$$p_Y(\omega_2) = \frac{2}{\pi} \sqrt{\frac{1}{1-\omega_2^2}}$$

Clearly then:

$$p_{X,Y}(\omega_1, \omega_2) \neq p_X(\omega_1) p_Y(\omega_2)$$

so $X \& Y$ are dependent.
Diagram showing a 3D graph with axes labeled x, y, and z. The point P(x, y) is indicated in red at the bottom left corner.
Conditional Independence

Maybe $x$ & $y$ are not independent, but that dependence is mediated by another r.v. $z$

E.g.:

```
x  ───>  z  ───>  y
  "intelligence"  "GPA"  "Salary"
```

Def: $x \perp y \mid z$ 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 $y \in \mathcal{Y} \doteq \{ \pm 1 \}$

Ex: spam classification of emails

$y = \text{spam} / \text{not spam}$

$x = \text{features computed from the email} = (\text{domain of sender}, \text{length of emails}, \text{entropy}, \text{counts of certain n-grams})$

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

$\hat{y} = \arg \max_{y \in \mathcal{Y} \doteq \{ \pm 1 \}} p(y|x)$
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)

Naïve Bayes assumption
Assume class-conditional independence of the features:

$$p(x|y) = \prod_{i=1}^{d} p(x_i|y)$$
From the NB assumption,

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

\[ = p(y=c) \frac{\prod_{i=1}^{d} p(x_i|y=c)}{\sum_{l=0}^{z+1} p(x, y=l)} \]

\[ = p(y=c) \frac{\prod_{i=1}^{d} p(x_i|y=c)}{p(y=-1)p(x|y=-1) + p(y=1)p(x|y=1)} \]
\[
\begin{align*}
p(y=c|x) &= p(y=c) \frac{\prod_{i=1}^{d} p(x_i|y=c)}{p(y=-1) \prod_{i=1}^{d} p(x_i|y=-1) + p(y=1) \prod_{i=1}^{d} p(x_i|y=1)} \\
p(y=-1) \prod_{i=1}^{d} p(x_i|y=-1) + p(y=1) \prod_{i=1}^{d} p(x_i|y=1)
\end{align*}
\]

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 \( i=1, \ldots, d \) — class-conditional probability of the \( i \)th observed feature

Can estimate these from the frequencies of the features in data.
Endnote: Computing marginal when \((x, y)\) uniform 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

\[
\int_{A} P_{x,y}(\omega_1, \omega_2) = 0 \quad \text{for any set } A \Rightarrow P(\mathbb{R}) = 0
\]

\[
\text{which is a contradiction}
\]

\[
\text{since } P_{x,y} \text{ is non-zero only on } \text{"a set of measure zero"}
\]

Really what we mean is

\[
(*) \quad P_{x,y}(\omega_1, \omega_2) = \frac{1}{2\pi} \delta(\omega_1^2 + \omega_2^2 - 1)
\]

where \(\delta\) is a special "function" (technically, a tempered distribution)

that satisfies

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

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,

\[ F(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t) \, dt \]

This follows from the fundamental theorem of calculus:

\[ \frac{d}{dx} F_X(x) = \frac{d}{dx} P(X \leq x) = \frac{d}{dx} \int_{-\infty}^{x} f_X(t) \, dt = f_X(x). \]

In our case, we see that

\[ F_X(x) = P(X \leq x) = P(C_x) \]

where \( C_x = \{ \omega_1, \omega_2 \} \) : \( \omega_1 \leq x \) and \( \omega_1^2 + \omega_2^2 = 1 \).
and $P(C_x)$ is the arc-length of $C_x$ divided by $2\pi$.

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,

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

so

$$P_x(x) = \frac{1}{\pi} \sqrt{\frac{1}{1-x^2}}.$$