Lecture 2  ML & Opt

- Recap of some probability
- Parametrized machine learning models
- Maximum likelihood estimation
Last lecture we saw that we model our ML problem as learning

\[ f: x \rightarrow y \]

\[ \text{targets} \]

\[ \text{predictor vars/features} \]

by modeling \( (x, y) \sim \mathcal{P} \) and minimizing the average loss of \( f \) w.r.t. \( \mathcal{P} \)
Distributions

$X \sim \mathcal{IP}$ means $\mathcal{IP}(X \in A)$ is the probability that $X$ takes on values in the set $A$

Two cases

$X$ is discrete:

$\mathcal{IP}(X \in A) = \sum_{x \in A} p_x(x)$ where $p_x$ is the probability mass function (pmf)

We have $\sum_{x} p_x(x) = 1$ and $p_x \geq 0$ everywhere

$X$ is continuous:

$\mathcal{IP}(X \in A) = \int_{A} p_x(x) dx$ where $p_x$ is the probability density function (pdf)

We have $\int_{\mathbb{R}} p_x(x) dx = 1$ and $p_x \geq 0$ everywhere
Examples

\( X \sim \text{Bern}(p) \) — Bernoulli distribution w/ parameter \( p \)
\[ p \in [0, 1] \]
means
\[ X = \begin{cases} 
1 & \text{with prob } p \\
0 & \text{w.p. } 1-p 
\end{cases} \]

\( X \sim \text{Binom}(n, p) \) — Binomial dist. w/ params \( n \) & \( p \)
means
\[ P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k \in \{0, 1, \ldots, n\} \]
$X \sim \text{Poisson}(\lambda) - \text{Poisson dist w/ rate parameter } \lambda$

$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$

$X \sim \mathcal{N}(\mu, \sigma^2) - \text{Normal or Gaussian r.v. w/ mean } \mu \text{ and variance } \sigma^2$

$p_X(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2}$
To model dependencies between random variables we use joint distributions

\[ P(\{x \in A, y \in B\}) \] written as \((x, y) \sim P\)

Analogously we have joint pdfs and joint pmfs. We assume there exists a joint pdf

\[ P(\{x \in A, y \in B\}) = \int_{A \times B} p_{xy}(x, y) \, dx \, dy \]

where \(p_{xy} \geq 0\) and

\[ \int_{\mathbb{R} \times \mathbb{R}} p_{xy} \, dx \, dy = 1 \]
Marginals

Given $p_{xy}$ we can extract the probability density of $X$

$$P(X \in A) = P(X \in A, Y \in \mathbb{R}) = \int_{A \times \mathbb{R}} p_{xy}(x, y) \, dx \, dy$$

$$= \int_{A} \left( \int_{\mathbb{R}} p_{xy}(x, y) \, dy \right) \, dx$$

marginal pdf of $x$

$$p_{x}(x) = \int_{\mathbb{R}} p_{xy}(x, y) \, dy$$

$p_{x}(x) > 0$ and $\int_{\mathbb{R}} p_{x}(x) \, dx = 1$
Similarly we get \( p_y \) by marginalizing over \( x \):

\[
p_y(y) = \int_{\mathbb{R}} p_{x,y}(x > y) \, dx
\]
In the best case, knowledge of $y$ conditional on $x$ is captured by the conditional distribution of $y$ given $x$

$$P_{y|x=x}(y) = P_y(y|X=x) = \frac{p_{xy}(x,y)}{p_x(x)}$$
Expectation

Average value of a r.v. \( X \)

\[
\mathbb{E}X = \int_{\mathbb{R}} x \cdot p(x) \, dx
\]

\[
\mathbb{E}X \quad X \sim \text{Bern}(p) \Rightarrow \mathbb{E}X = 0 \cdot (1-p) + 1 \cdot p = p
\]

\[
X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \mathbb{E}X = \mu
\]
Expectation of random vectors

If \( \{ \xi \} \sim P \) then \( \mathbb{E} [\xi] = \int_{\mathbb{R} \times \mathbb{R}} [\xi] \, p_{\xi \gamma}(x, y) \, dx \, dy \)

\[
= \int_{\mathbb{R} \times \mathbb{R}} \left[ x \, p_{\xi \gamma}(x, y) \right] \, dx \, dy
\]

\[
= \left[ \int_{\mathbb{R}} x \, p_{\xi \gamma}(x, y) \, dx \right] \, dy
\]

\[
= \left[ \mathbb{E}_x \left[ \xi \right] \right] = \left[ \mathbb{E}_X \right]
\]
**Conditional Expectation**

\[ E(Y | X) \] — average value of \( Y \) given knowledge of \( X \)

\[ E(Y | X=x) = \int y \rho_{Y|X}(y | x=x) \, dy \]
Variance

Variance measures how close a random variable stages to its mean, on average.

\[ \text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 \]

\[ = \int_{\mathbb{R}} (x - \mathbb{E}X)^2 \rho_X(x) \, dx \]

\[ = \int_{\mathbb{R}} [x^2 - 2x \mathbb{E}X + (\mathbb{E}X)^2] \rho_X(x) \, dx \]

\[ = \int_{\mathbb{R}} x^2 \rho_X(x) \, dx - 2 \mathbb{E}X \int_{\mathbb{R}} x \rho_X(x) \, dx + (\mathbb{E}X)^2 \]

\[ = \mathbb{E}(X^2) - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 \]
Ex of Variance

$X \sim \text{Bern}(\rho)$

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

$= \mathbb{E}X - (\mathbb{E}X)^2$

$= \rho - \rho^2$

Variances is a measure of uncertainty

$X \sim \mathcal{N}(\mu, \sigma^2)$

$\sigma^2 > \sigma^2$
Independence \( (X \perp Y) \)

We say two random variables are independent if

\[
P(x \in A, y \in B) = P(x \in A) P(y \in B)
\]

\[
P_{X,Y}(x,y) = P_X(x) P_Y(y)
\]

\[
P_Y(y | x = x) = P_Y(y) \quad \text{and} \quad P_X(x | y = y) = P_X(x)
\]
Conditional Independence

We say $X \notin Y$ are conditionally independent given $Z$ if

$$P_{X \rightarrow Y | Z} = P_{X | Z} \cdot P_{Y | Z}$$

$$P_{X \rightarrow Y | Z = z}^{(x, y)} = \frac{P_{X \rightarrow Y \rightarrow Z}^{(x, y, z)}}{P_z(z)} \text{ in general}$$

$$= P_{X | Z = z(x)} \cdot P_{Y | Z = z(y)}$$

if $X \perp Y | Z$

Ex: SAT(A) → Academic Ability (Z) → GRE(B) → ALL B | Z
**Parameterized ML models**

\[ p(y | x) \] -- captures everything about values of \( y \) given knowledge \( X = x \)

\[ E[y | x = x] \] -- captures a point estimate of \( y \) given \( x \)

Issues w/ using these in ML: either can be an arbitrarily complicated unknown function of \( x \).

To deal with this, we use **parameterized models**,

i.e.

\[ p(y | x) = f_{\theta}(x) \] where \( \theta \in \mathbb{R}^p \) is a parameter vector

\[ = f(x; \theta) \]
Ex

**OLS:** \[ y = \beta^T x + \varepsilon \] where \( \varepsilon \sim \mathcal{N}(0, \frac{1}{n}) \)

i.e. \( y|x \sim \mathcal{N}(\beta^T x, \frac{1}{n}) \)

Assume data \((x_i, y_i)\) came from this type of model for \( y|x \), then we learn \( \beta \) that fits this data best.

OLS is appropriate for data that looks like

and inappropriate for, e.g.
what about parameterized models for data that doesn't fit the OLS assumptions?

e.g.
- $y$ is not a cont. r.v.
- $y$ has a limited range
- $y - \mathbb{E}[y|x]$ is not indep of $x$
- Poisson/"shot noise" model corresponds to $\text{Poisson}(\lambda)$ being used, where $\lambda$ is a function of $x$

$$y \mid x \sim \text{Poisson}(\lambda(x))$$

e.g. this gives

$$\mathbb{E}[y \mid x] = \lambda_\theta(x)$$

Notice since $\lambda_\theta(x)$ is a rate, it must satisfy $\lambda_\theta(x) > 0$. We impose this by taking

$$\lambda_\theta(x) = e^{f(x; \theta)}$$