Lecture 1

Sunday, January 21, 2024 5:29 AM

Probability Review

Reminder: HW due Thurs.

Readings: Murphy §2, §3 ← I'll follow this, but adopt standard probability/set notation
Bishop §2

Note: I will update these notes as course progresses to maintain probability you will need to know

For extra resources filling in gaps in background

- "Probability Essentials" Jacod + Protter
  - Short book w/ self-contained chapters + exercises - good for self-study
- "Probability: theory + examples" Durrett
  - Measure theoretic + rigorous coverage for those interested in serious research on probability + ML

- Questions on HW? (Kolmogorov)

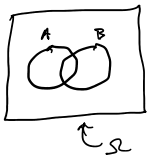
Def Probability space is a triple $(\Omega, \mathcal{F}(\Omega), \mathbb{P})$ consisting of:

- Ω - a sample space
  all events which may occur
  ex Flipping two coins
  $\Omega = \{H_1, H_2, T_1, T_2\}$
- $\mathcal{F}$ - an event space, typically taken as the "σ-algebra" consisting of countably finite intersections & unions of elements from Ω
  ex the event that both coins are heads $H_1 \cap H_2$
- $\mathbb{P}$ - a probability measure
  $0 \le \mathbb{P}(A) \le 1, \quad A \in \mathcal{F}$

↳ frequentist perspective
Given $N \to \infty$ identical experiments, the proportion where the event happens

↳ Bayesian perspective
A "modeling function" describing how likely an event is to occur

- Probability of two events



def Given $A, B \in \mathcal{F}$
AND $\mathbb{P}(A, B) := \mathbb{P}(A \cap B)$
OR $\mathbb{P}(A \cup B)$

def Set operations + defs

- Set subtraction $A \setminus B$ or $A - B = A \cap B^c$
- deMorgan's Laws $(\cup A_i)^c = \cap A_i^c$
  $(\cap A_i)^c = \cup A_i^c$
- Distributive law $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- Cartesian product $A \times B = \{(x, y) \mid x \in A, \ y \in B\}$
- cardinality $\#A$ = number of elements in A
- inclusion/exclusion principle $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
  or more generally

- inclusion/exclusion principle

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

or more generally

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \wedge A_j) - \cdots + (-1)^{n+1} \mathbb{P}(A_1 \wedge \cdots \wedge A_n)$$

exercise Prove that $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$

$$\mathbb{P}(B) = \mathbb{P}(A \cup (B \setminus A))$$
$$= \mathbb{P}(A) + \mathbb{P}(B \setminus A) \quad \leftarrow \text{no intersection}$$
$$\geq \mathbb{P}(A) \qquad \mathbb{P}(B \setminus A) \geq 0$$

Def Conditional Probability & Dependence

- The probability of B conditioned on A

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A,B)}{\mathbb{P}(A)}$$

- Events A & B are independent if

$$\mathbb{P}(A,B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

i.e. $\mathbb{P}(B|A) = \mathbb{P}(B)$ (Knowledge of A tells you nothing about B)

→ we write $A \perp B$ in shorthand

- Events $A_1, \ldots A_N$ are conditionally indep. on B if

$$\mathbb{P}(A_1, \ldots, A_N | B) = \prod_{i=1}^{N} \mathbb{P}(A_i | B)$$

Random Variables

A function which takes a value w/ a given probability

def a RV taking a finite set of values $X = \{x_1, \ldots x_N\}$ is a discrete RV

sample/state space

We denote $p(x_i) = \mathbb{P}(X = x_i)$ as the probability mass function, noting

① $0 \leq p(x) \leq 1$

② $\sum_{x \in X} p(x) = 1$

Notation: sum over all x in X

def a RV X taking values over $\mathbb{R}$ is a continuous RV. We describe prob. over intervals:

- $\mathbb{P}(a \leq x \leq b) = \mathbb{P}(x \leq b) - \mathbb{P}(x \leq a)$
- we define $F_X(x) = \mathbb{P}(X \leq x)$ as cumulative distribution function (CDF)

CDF/PDF

- If F is differentiable we can work w/ the probability density function (PDF)

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Notation: sometimes we drop the X subscript

Allowing us to work w/ either

$$\mathbb{P}(a \leq x \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

Sets of RVs

Joint distribution $f(x_1, \ldots x_N) = \mathbb{P}(X_1 = x_1, \ldots X_N = x_N)$

Marginal distribution $f(X = x) = \sum_y \mathbb{P}(X = x, Y = y)$

AKA rule of total probability

Conditional distribution $f(Y = y | X = x) = \frac{f(X = x, Y = y)}{f(X = x)}$

conditional distribution $f(Y=y \mid X=x) = \dfrac{f(X=x, Y=y)}{f(X=x)}$ (probability)

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

probabilistic chain rule
$$f(x_1, \dots x_N) = f(x_2, \dots x_N | x_1)\, f(x_1)$$
$$\vdots = f(x_3 \dots x_N | x_1, x_2)\, f(x_2|x_1)\, f(x_1)$$
$$= f(x_N | x_1, \dots x_N) \cdots f(x_2|x_1)\, f(x_1)$$

marginal independence $X \perp Y \iff f(x,y) = f(x)\, f(y)$ or more generally
$$f(x_1, \dots x_N) = \prod_{i=1}^{N} f(x_i)$$

conditional independence $X \perp Y \perp Z \iff f(x,y|z) = f(x|z)\, f(y|z)$

Expectations, Variances + other Moments

def $\mathbb{E}_X[X] := \int x\, f_X(x)\, dx$ or $\sum_{x \in \mathcal{X}} x\, p(x)$ expectation

for cont. RV for discrete RV

can drop if clear

linearity $\mathbb{E}[\alpha X + \beta] = \alpha\, \mathbb{E}[X] + \beta$

indep. $\mathbb{E}_{x_1 \dots x_N}\left[\prod_{i=1}^{N} X_i\right] = \prod_{i=1}^{N} \mathbb{E}_{x_i}[X_i]$ if x_i are indep

def $\mathrm{Var}(X) := \mathbb{E}_x\left[(x - \underbrace{\mathbb{E}_x(x)}_{\mu})^2\right] = \sigma^2$

equivalently $= \mathbb{E}_x[x^2] - \mu^2$

so $\mathbb{E}[x^2] = \mu^2 + \sigma^2$

linearity $\mathrm{var}[\alpha X + \beta] = \alpha^2\, \mathrm{var}[X]$

Law of total expectation $\mathbb{E}_x[X] = \mathbb{E}_Y\left[\mathbb{E}_x[X|Y]\right]$

Law of total variance $\mathrm{var}[X] = \mathbb{E}\left[\mathrm{var}[X|Y]\right] + \mathrm{var}\left[\mathbb{E}[X|Y]\right]$

Bayes rule

From definition of conditional distribution

$$\mathbb{P}(x,y) = \mathbb{P}(x|y)\,\mathbb{P}(y) = \mathbb{P}(y|x)\,\mathbb{P}(x)$$

$$\Rightarrow \mathbb{P}(y|x) = \frac{\mathbb{P}(x|y)\,\mathbb{P}(y)}{\mathbb{P}(x)}$$

- $\mathbb{P}(y|x)$: posterior distribution
- $\mathbb{P}(x|y)$: likelihood
- $\mathbb{P}(y)$: prior — knowledge of y w/o access to x
- $\mathbb{P}(x)$: marginal likelihood

think of x as hidden data directly we can infer

we can infer

marginal

w/o access to x

Probability Distributions

Lots to cover → we'll focus on two that form the cornerstone of classification and regression

Bernoulli + binomial dist

Probability of coin landing heads given by $0 \leq \theta \leq 1$

$$Y \sim \text{Ber}(\theta)$$

$$P(Y) = \begin{cases} 1-\theta & Y = 0 \\ \theta & Y = 1 \end{cases}$$

$$= \theta^Y (1-\theta)^Y$$

Given N experiments, how many are heads?

$$S = \sum_{i=1}^{N} Y$$

$$S \sim \text{Bin}(N, \theta)$$

$$P(S) = \binom{N}{S} \theta^S (1-\theta)^{N-S}$$

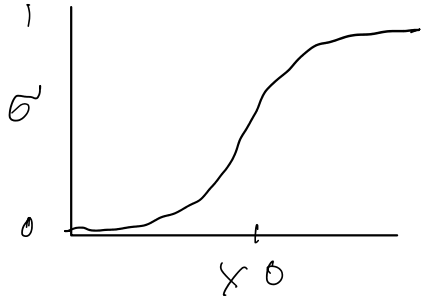
$\binom{N}{K}$ is "N choose K" $= \frac{N!}{(N-K)!\,K!}$

Sigmoid Softmax

$$\sigma(x) = \frac{1}{1+e^{-a}}$$

$$\sigma'(x) = \frac{d}{dx}\sigma_+(x)$$

$$\sigma_+(x) = \log(1+e^x)$$



$a \to \infty$ recovers heaviside

Binary Logistic Regression

Given set of data we would like to perform binary classification

$$p(y \mid x, \theta) = \text{Ber}(y \mid \sigma(w^T x + b))$$

$$p(y \mid x, \theta) = \text{Ber}(y \mid \sigma(w \cdot x + b))$$

★ Differentiable functions may be chained together and fit w/ gradient descent (HW1)

Normal Distributions

$$Y \sim \mathcal{N}(y; \mu, \sigma^2)$$

iff

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) \quad \text{PDF}$$

$$F(y) = \int_{-\infty}^{y} f(y')\, dy' \quad \text{CDF}$$

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} y f(y)\, dy$$

$$= \mu$$

$$\text{var}(Y) = \mathbb{E}[Y^2] - \mu^2$$

$$\mathbb{E}[Y^2] = \sigma^2 + \mu^2$$

Hetero/homo skedastic Regression

$$p(y; x, \theta) = \mathcal{N}(y \mid f_\mu(x;\theta), f_\sigma(x;\theta))$$

$f_\mu \in \mathbb{R}$

$f_\sigma \in \mathbb{R}^+$ ← need to enforce to get valid

homo skedastic $\rightarrow$ $\sigma^2 \neq f(x)$

hetero skedastic $\rightarrow$ $\sigma^2 = f(x)$

$$p(y \mid x, \theta) = \mathcal{N}(y \mid w_\mu^T x + b_\mu, \sigma_+(w_\sigma^T x))$$

Other Distributions

Make sure you are familiar w/ other dist (§2.7 in Murphy)