THEORY
OF
PROBABILITY
&
ESTIMATION

This page intentionally left blank

credits : A. Moore
Tom Mitchell

Random Variable

- Outcome of a randomized experiment
- denotes something about which we are uncertain

e.g.

A = True if a randomly drawn person is Female

define P(A) as fraction of possible worlds in which A is true" in repeated runs of the random experiment

- set of possible world s is called sample space S
- random variable A is a function defined over S

$$A : S \rightarrow \{0, 1\}$$

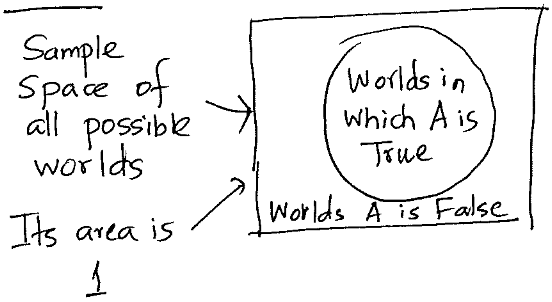
Event

- a subset of S

e.g. - subset of S for which gender = f
- subset of S for which (gender = m) AND (eyeColor = Blue)

we are interested in propabilities of events or events conditioned on others

Visualizing P(A) :



P(A) = Area of worlds A is True

The Axioms of Probability

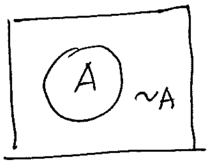
- $0 \leqslant P(A) \leqslant 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = \emptyset$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

[di Finetti 1931] Not Using these axioms when Gambling → opponent can exploit you!

Corollary

- $P(\sim A) + P(A) = 1$

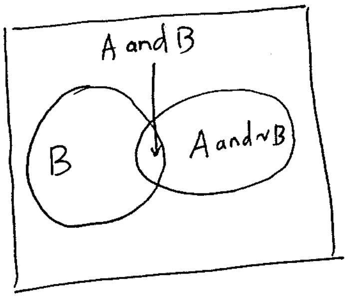
Proof Sketch:



- $P(A) = P(A \text{ and } B) + P(A \text{ and } \sim B)$

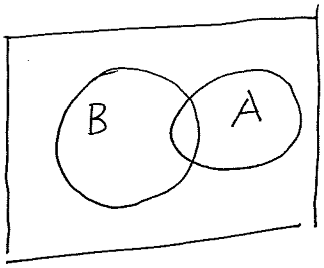
Proof Sketch:

$$A = [A \text{ and } (B \text{ or } \sim B)]$$
$$= [(A \text{ and } B) \text{ or } (A \text{ and } \sim B)]$$
$$P(A) = P([A \text{ and } B) \text{ or } (A \text{ and } \sim B))$$
$$= P(A \text{ and } B) + P(A \text{ and} \sim B) - P((A \text{ and } B) \text{ and } (A \text{ and } \sim B))$$
$$= P(A \text{ and } B) + P(A \text{ and } \sim B)$$



Conditional Probability

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



Corollary: The chain Rule

$$P(A \text{ and } B) = P(A|B)\, P(B)$$
$$P(A \text{ and } B \text{ and } C) = P(A|BC)\, P(BC)$$
$$= P(A|BC)\, P(B|C)\, P(C)$$

Bayes Rule

$$P(A B) = P(A|B)\, P(B)$$
$$= P(B|A)\, P(A)$$

$$\Rightarrow P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}$$ Bayes' rule (1763)

Thomas Bayes

P(A) is called "prior"

P(A|B) is called "posterior"

$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B|A)\, P(A) + P(B|\sim A)\, P(\sim A)} \; ; \quad P(A|B X) = \frac{P(B|AX) P(AX)}{P(BX)}$$

Applyling Bayes Rule

A = you have the flu, B = you just coughed

Assume :

$$P(A) = 0.05$$
$$P(B|A) = 0.80$$
$$P(B|\sim A) = 0.20$$

what is $P(\text{flu} \mid \text{cough}) = P(A|B)$

$$= \frac{P(B/A)\ P(A)}{P(B|A)\ P(A) + P(B|\sim A)\ P(\sim A)}$$

$$= \frac{0.80 * 0.05}{0.80 * 0.05 + 0.20 * 0.95} = 0.17$$

Function Approximation

insted of $f : X \to Y$

learn $P(Y|X)$

Joint Distribution

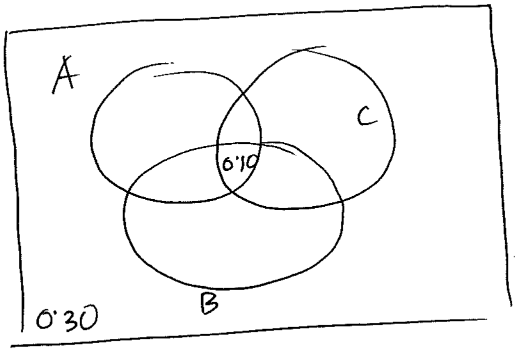
Steps for making a joint distribution of M variables:

Step1 : Make a truth table listing all combinations (M r.v → 2^M rows)

Step2 : For each combination of values, say how likely it is.

Step3 : By axioms of prob. those numbers must sum to 1.

A	B	C	Prob.
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint Distribution

Once you have the JD, you can ask for the probability of any logical expressions involving these variables

This sounds like the solution to learning $f : X \to Y$ or $P(Y|X)$

Are we done ?

Main Problem: learning $P(Y|X)$ can require more data than we have

consider learning JD with 100 attributes

$\#$ rows in this table ? $2^{100} \sim 1000^{10} = 10^{30}$ $2^{10} = 1024$

$\#$ people on earth ? ~billion 10^9

fraction of rows with ϕ training examples ? ~ 0.9999

- Huge problem in Data Science
- Data sparcity !

What to do?

1. Be smart about estimating probabilities from sparse data
   - maximum likelihood estimates
   - maximum a posteriori estimates
2. Be smart about how to represent joint distributions
   - Bayes nets, graphical models

credits: A. Moore
T. Mitchell

PART I: Be smart about how we estimate probabilities

Estimating Probability of Heads

- Given a coin, estimate prob. that it will turn up heads $(X=1)$ or tails $(X=0)$
- You flip repeatedly, observing
  - it turns up heads α_1 times
  - it turns up tails α_2 times
- What would be your estimate for $P(X=1)$? $\frac{\alpha_1}{\alpha_1+\alpha_0} = \hat{P}(X=1)$

Test A: 100 flips, $\overset{\alpha_1}{51}$ heads $(X=1)$, $\overset{\alpha_0}{49}$ tails $(X=0)$

$$\frac{\alpha_1}{\alpha_1+\alpha_0} = \frac{51}{100} = 0.51 \leftarrow \hat{P}(X=1)$$

Test B: 3 flips, $\overset{\alpha_1}{2}$ heads $(X=1)$, $\overset{\alpha_0}{1}$ tail. $(X=0)$

$$\frac{\alpha_1}{\alpha_1+\alpha_0} = \frac{2}{2+1} = 0.666$$

Test C: Keep flipping, and develop a single learning algorithm (online learning) that gives reasonable estimate after each flip.

$$\frac{\alpha_1+\beta_1}{(\alpha_1+\beta_1)+(\alpha_0+\beta_0)}$$

β_1 = # of hallusinated flips that tumes up heads

β_0 = # of hallusinated flips that turned up tails

The number of hallusinated flips and their outcomes are "priors"

Stronger the "prior", more actual data will be needed to converge to the observed ground truth.

Principles for Estimating Probabilities

- Principle 1 (maximum likelihood):
  - choose parameters θ that maximize $P(\text{data} \mid \theta)$

e.g. $$\hat{\theta}^{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

- Principle 2 (maximum a posteriori):
  - choose parameters θ that maximize $P(\theta \mid \text{data})$

e.g. $$\hat{\theta}^{MAP} = \frac{\alpha_1 + \#\text{ hallucinated\_1s}}{(\alpha_1 + \#\text{ hallucinated\_1s}) + (\alpha_0 + \#\text{ hallucinated\_0s})}$$

$$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta)\, P(\theta)}{P(\text{data})} \qquad \text{Bayes Rule}$$

Formal Treatment: Maximum Likelihood Estimation (Principle 1)

$$P(X=1) = \theta \qquad P(X=0) = (1-\theta)$$

Data D : 1 0 0 1 1

$$P(D \mid \theta) = \theta \quad (1-\theta) \quad (1-\theta) \quad \theta \quad \theta \qquad \theta^{\alpha_1} (1-\theta)^{\alpha_0}$$

Flips produce data D with α_1 heads, α_0 tails, iid ~ Bernoulli

α_1 and α_0 are counts that sum these outcomes (Binomial)

- $$\hat{\theta} = \arg\max_{\theta} \ln P(D \mid \theta) = \arg\max_{\theta} \ln \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

- Set derivate to zero $$\frac{d}{d\theta} \ln P(D \mid \theta) = 0$$

$$\hat{\theta} = \arg\max_{\theta} \ln P(D|\theta) = \arg\max_{\theta} \ln\left[\theta^{\alpha_1}(1-\theta)^{\alpha_0}\right]$$

$$= \arg\max_{\theta} \left[\alpha_1 \ln\theta + \alpha_0 \ln(1-\theta)\right]$$

$$\frac{\partial}{\partial\theta}\alpha_1 \ln\theta + \frac{\partial}{\partial\theta}\alpha_0 \ln(1-\theta) \qquad \frac{\partial \ln\theta}{\partial\theta} = \frac{1}{\theta}$$

$$= \alpha_1 \frac{1}{\theta} + \alpha_0 \frac{\partial \ln(1-\theta)}{\partial\theta}$$

$$= \alpha_1 \frac{1}{\theta} + \alpha_0 \frac{\partial \ln(1-\theta)}{\partial(1-\theta)} \cdot \frac{\partial(1-\theta)}{\partial\theta} \qquad \text{chain rule}$$

$$= \alpha_1 \cdot \frac{1}{\theta} + \alpha_0 \cdot \frac{1}{1-\theta} \cdot (-1)$$

$$\frac{\alpha_1}{\theta} - \alpha_0 \frac{1}{1-\theta} = 0 \implies \boxed{\hat{\theta}^{MAP} = \frac{\alpha_1}{\alpha_1+\alpha_0}}$$

Summary,

- $X \sim$ Bernoulli : $P(X) = \theta^{X}(1-\theta)^{(1-X)}$
- Data set D of iid produces α_1 ones, α_0 zeros (Binomial)

  $P(D|\theta) = P(\alpha_1, \alpha_0 | \theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$

  $\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta) = \dfrac{\alpha_1}{\alpha_1+\alpha_0}$

Formal Treatment : maximum a posteriori (MAP) Estimation (Principle 2)

Beta Prior distribution

Prior : $P(\theta) = \dfrac{\theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$

Likelihood : $P(D|\theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$

Posterior : $P(\theta|D) \propto P(D|\theta) P(\theta)$

$$P(D|\theta) P(\theta) = \frac{\theta^{\alpha_H + \beta_H - 1} (1-\theta)^{\alpha_T + \beta_T - 1}}{B(\beta_H, \beta_T)}$$

$$\boxed{\hat{\theta}^{MAP} = \frac{(\alpha_H + \beta_H - 1)}{(\alpha_H + \beta_H - 1) + (\alpha_T + \beta_T - 1)}}$$

- Role of Beta distribution is to play the role of hallucinated flips.
- This in infact called conjugate prior

Summary , likelihood is $\sim$ Binomial $P(D|\theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}}{B(\beta_H . \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

Then, posterior is Beta distribution $P(\theta|D) \approx \text{Beta}(\alpha_H + \beta_H, \alpha_T + \beta_T)$

and MAP estimate is

$$\hat{\theta}^{MAD} = \frac{(\alpha_H + \beta_H - 1)}{(\alpha_H + \beta_H - 1) + (\alpha_T + \beta_T - 1)}$$

Example of Dice Roll (6 outcomes instead of 2)

Likelihood is $\sim$ Multinomial $\left(\theta = \{\theta_1, \theta_2, \cdots \theta_K\}\right)$

$$P(D|\theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \cdots \theta_K^{\alpha_K}$$

Prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1 - 1} \theta_2^{\beta_2 - 1} \cdots \theta_K^{\beta_K - 1}}{B(\beta_1, \beta_2, \cdots, \beta_K)} \sim \text{Dirichlet}(\beta_1, \cdots \beta_K)$$

Posterior is a Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \beta_2 + \alpha_2, \cdots, \beta_K + \alpha_K)$$

and MAP estimate is

$$\hat{\theta}^{MAD} = \frac{\alpha_i + \beta_i - 1}{\sum_{j=1}^{K} (\alpha_j + \beta_j - 1)}$$