Probability Review

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
    - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
    - Independence, conditional independence
- Mean and Variance
- The big picture: probabilistic inference
- Examples

Why probability?

- Probability is the proper mechanism for accounting for *uncertainty*
  - Measurement noise and incomplete knowledge
  - Is the image dark because the light level is low, or because the surface has low albedo? It is less common to see very dark surfaces under very bright lights than it is to see a range of albedoes under a reasonably bright light.
  - 2D to 3D (shape from X)
  - Variations: scale, rotation, deformation, occlusion

Sample space and Events

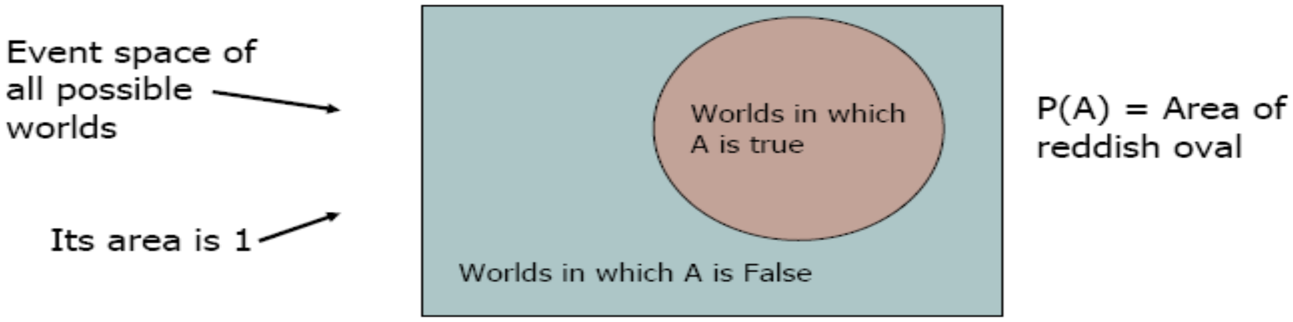
- Ω : Sample Space, result of an experiment
  - If you toss a coin twice $\Omega = \{\mathrm{HH},\mathrm{HT},\mathrm{TH},\mathrm{TT}\}$
- Event: a subset of Ω
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Closed under finite union and complements
    - Entails other binary operation: union, diff, negation, intersection, etc.
  - Contains the empty event and Ω

Probability Measure

- Defined over (Ω,S) s.t.
  - P(α) >= 0 for all α in S
  - P(Ω) = 1
  - If α, β are disjoint, then
    - P(α U β) = p(α) + p(β)
- We can deduce other axioms from the above ones
  - Ex: P(α U β) for non-disjoint event

    P(α U β) = p(α) + p(β) – p($\alpha \cap \beta$)

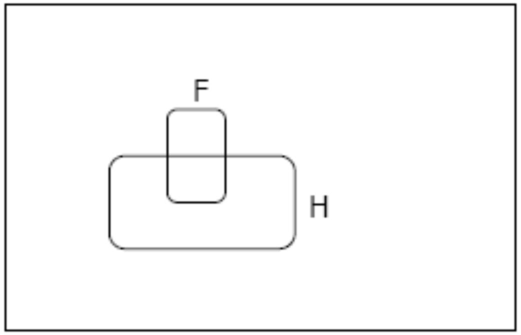
Visualization


Event space of all possible worlds
Its area is 1
Worlds in which A is true
Worlds in which A is False
P(A) = Area of reddish oval

- We can go on and define conditional probability, using the above visualization

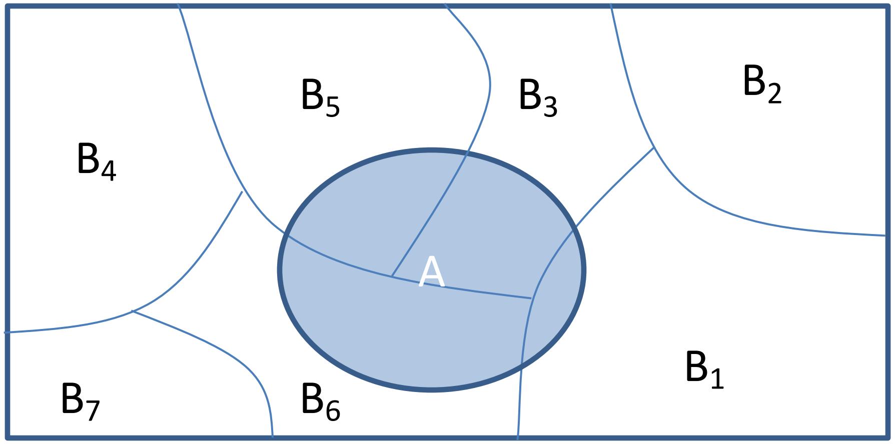
Conditional Probability

P(F|H) = Fraction of worlds in which H is true that also have F true



$$p(f \mid h) = \frac{p(F \cap H)}{p(H)}$$

Rule of total probability

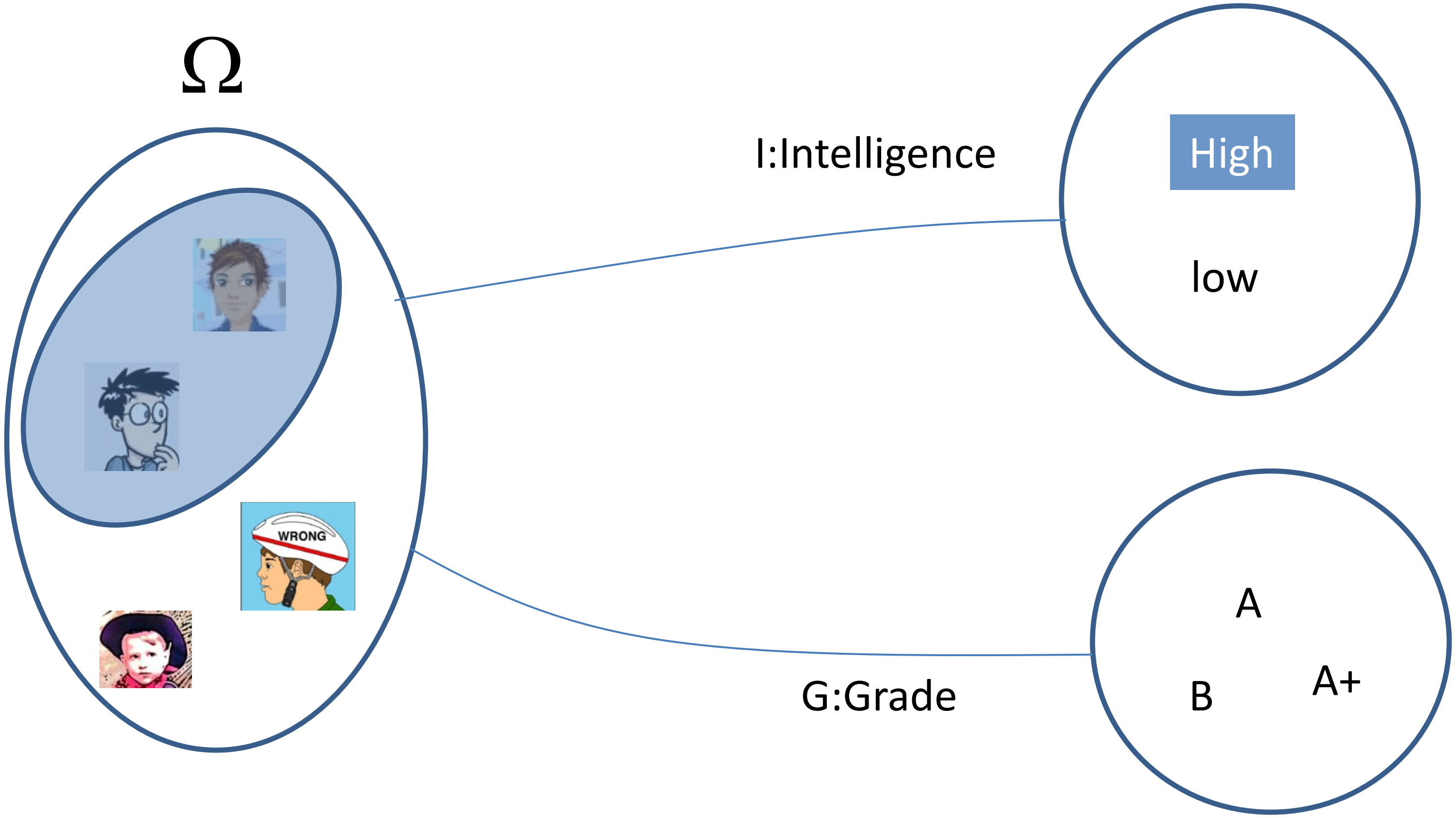


$$p(A) = \sum P(B_i)P(A \mid B_i)$$

From Events to Random Variable

- In many learning problems we will be dealing with RV
- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - $\Omega =$ all possible students
  - What are events?
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - Intelligence_High = … with high intelligence
  - Very cumbersome
  - We need “functions” that maps from Ω to an attribute space.
  - P(G = A) = P({student ϵ Ω : G(student) = A})

Random Variables



Ω

I:Intelligence

High

low

G:Grade

A

B

A+

P(I = high) = P({all students whose intelligence is high})

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of tails X you get if you flip 100 coins
- X is a RV with **arity** *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$
  - E.g. the possible values that X can take on are 0, 1, 2, …, 100

Probability of Discrete RV

- Probability mass function (pmf): $P(X = x_i)$
- Easy facts about pmf
  - $\Sigma_\mathrm{i}\ P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$

Common Distributions

- Uniform X *U*[1, …, *N*]
  - X takes values 1, 2, … *N*
  - P(X = *i*) = 1/*N*
  - E.g. picking balls of different colors from a box
- Binomial X *Bin*(*n*, *p*)
  - X takes values 0, 1, …, *n*
  - $p(X=i)=\binom{n}{i}p^{i}(1-p)^{n-i}$
  - E.g. coin flips, P(k heads and n − k tails in n flips)

Continuous Random Variables

- Probability *density* function (pdf) instead of probability *mass* function (pmf)
- A pdf is any function $f(x)$ that describes the probability density in terms of the input variable x.

Probability of Continuous RV

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$
- Actual probability can be obtained by taking the <u>integral</u> of pdf
  - E.g. the probability of X being between 0 and 1 is

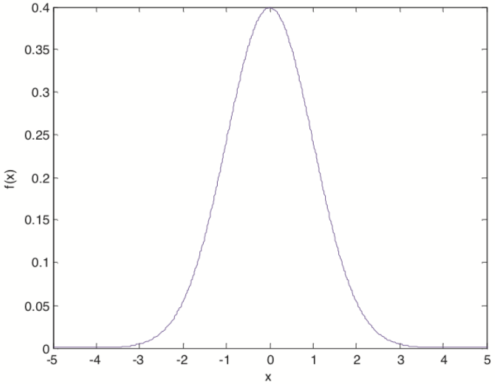
$$P(0 \leq X \leq 1) = \int_{0}^{1} f(x)dx$$

Cumulative Distribution Function

- $F_X(v) = P(X \leq v)$
- Discrete RVs
  - $F_X(v) = \Sigma_{vi}\ P(X = v_i)$
- Continuous RVs
  - $F_X(v) = \int_{-\infty}^{v} f(x)dx$
  - $\frac{d}{dx} F_x(x) = f(x)$

Common Distributions

- Normal X $N(\mu, \sigma^2)$
  - $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$
  - E.g. the height of the entire population
  - Grades of the entire class?



Multivariate Normal

- Generalization to higher dimensions of the one-dimensional normal

Covariance matrix

$$f_{\vec{X}}(x_i, \ldots, x_d) = \frac{1}{(2\pi)^{d/2} \left|\Sigma\right|^{1/2}}$$

$$\cdot \exp\left\{ -\frac{1}{2} (\vec{x} - \mu)^T \Sigma^{-1} (\vec{x} - \mu) \right\}$$

Mean (vector)

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Mean and Variance
- The big picture: probabilistic inference
- Examples

Joint Probability Distribution

- Random variables encodes attributes
- Not all possible combination of attributes are equally likely
    - Joint probability distributions quantify this
- P(X= x, Y= y) = P(x, y)
    - Generalizes to N-RVs
    - $\sum_x \sum_y P\left(X = x, Y = y\right) = 1$
    - $\int_x \int_y f_{X,Y}\left(x, y\right) dx dy = 1$

Chain Rule

- Always true
  - $P(x, y, z) = p(x)\, p(y|x)\, p(z|x, y)$

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

    $= \ldots$

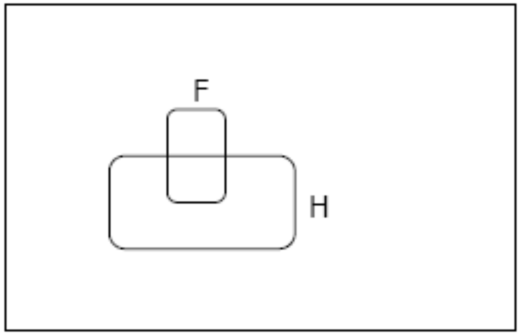
Conditional Probability

events

$$\mathrm{P}\left(\mathrm{X}=x \,\middle|\, \mathrm{Y}=y\right)=\frac{\mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y\right)}{\mathrm{P}\left(\mathrm{Y}=y\right)}$$

But we will always write it this way:

$$P\left(x \mid y\right)=\frac{p(x,y)}{p(y)}$$

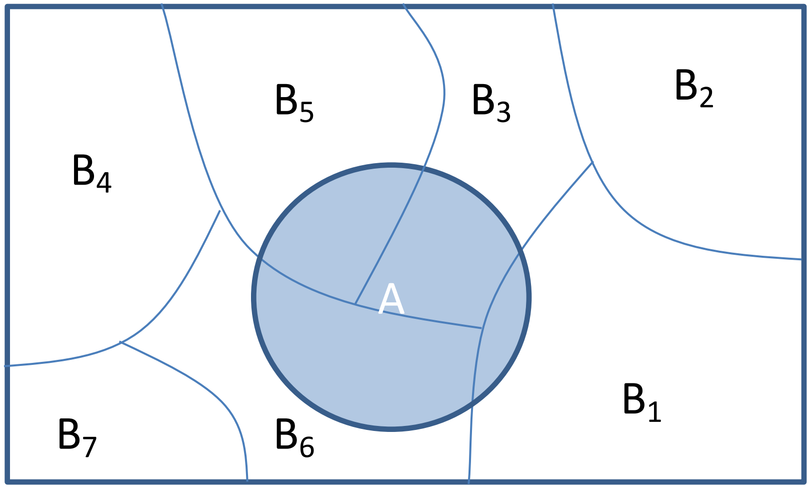


Marginalization

- We know p(X, Y), what is P(X=x)?
- We can use the law of total probability

$$p(x) = \sum_y P(x, y)$$

$$= \sum_y P(y) P(x \mid y)$$



Marginalization Cont.

- Another example

$$p(x) = \sum_{y,z} P(x, y, z)$$

$$= \sum_{z,y} P(y, z) P(x \mid y, z)$$

Bayes Rule

- We know that P(rain) = 0.5
  - If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$P\left(rain \mid wet\right) = \frac{P(rain)P(wet \mid rain)}{P(wet)}$$

$$P\left(x \mid y\right) = \frac{P(x)P(y \mid x)}{P(y)}$$

Bayes Rule cont.

- You can condition on more variables

$$P\left(x \mid y,z\right)=\frac{P(x \mid z)P(y \mid x,z)}{P(y \mid z)}$$

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
    - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
    - Independence, conditional independence
- Mean and Variance
- The big picture: probabilistic inference
- Examples

Independence

- X is independent of Y means that knowing Y does not change our belief about X.
  - P(X|Y=y) = P(X)
  - P(X=x, Y=y) = P(X=x) P(Y=y)
  - The above should hold for all x, y
  - It is symmetric and written as $X \perp Y$

Independence

- $X_1, \ldots, X_n$ are independent <u>if and only if</u>

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P\left(X_i \in A_i\right)$$

- If $X_1, \ldots, X_n$ are independent and identically distributed we say they are *iid* (or that they are a random sample) and we write

$$X_1, \ldots, X_n \sim P$$

CI: Conditional Independence

- RV are rarely independent but we can still leverage local structural properties like Conditional Independence.
- $X \perp Y \mid Z$ if once Z is observed, knowing the value of Y does not change our belief about X
  - P(rain $\perp$ sprinkler's on | cloudy)
  - P(rain $\not\perp$ sprinkler's on | wet grass)

Conditional Independence

- $P(X{=}x \mid Z{=}z, Y{=}y) = P(X{=}x \mid Z{=}z)$
- $P(Y{=}y \mid Z{=}z, X{=}x) = P(Y{=}y \mid Z{=}z)$
- $P(X{=}x, Y{=}y \mid Z{=}z) = P(X{=}x \mid Z{=}z)\, P(Y{=}y \mid Z{=}z)$

We call these **factors** : very useful concept !!

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Mean and Variance
- The big picture: probabilistic inference
- Examples

Mean and Variance

- Mean (Expectation): $\mu = E(\mathrm{X})$
  - Discrete RVs: $E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$

$$E(g(X)) = \sum_{v_i} g(v_i) P(X = v_i)$$

  - Continuous RVs: $E(\mathrm{X}) = \int_{-\infty}^{+\infty} x f(x)\, dx$

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

Mean and Variance

- Variance: $Var(X) = E((X-\mu)^2)$

  $$Var(X) = E(X^2) - \mu^2$$

  - Discrete RVs: $V(\mathrm{X}) = \sum_{v_i} (v_i - \mu)^2 \, \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $V(\mathrm{X}) = \int_{-\infty}^{+\infty} (x-\mu)^2 \, f(x) dx$

- Covariance:

$$Cov(X,Y) = E((X-\mu_x)(Y-\mu_y)) = E(XY) - \mu_x\mu_y$$

Mean and Variance

- Correlation:

$$\rho(X,Y) = Cov(X,Y)/\sigma_x\sigma_y$$

$$-1 \leq \rho(X,Y) \leq 1$$

Properties

- Mean
  - $E\left(\mathrm{X}+\mathrm{Y}\right)=E\left(\mathrm{X}\right)+E\left(\mathrm{Y}\right)$
  - $E\left(a\mathrm{X}\right)=aE\left(\mathrm{X}\right)$
  - If X and Y are independent, $E\left(\mathrm{XY}\right)=E\left(\mathrm{X}\right)\cdot E\left(\mathrm{Y}\right)$
- Variance
  - $V\left(a\mathrm{X}+b\right)=a^2V\left(\mathrm{X}\right)$
  - If X and Y are independent, $V\left(\mathrm{X}+\mathrm{Y}\right)=V(\mathrm{X})+V(\mathrm{Y})$

Some more properties

- The conditional expectation of Y given X when the value of X = x is:

$$E\left(Y \mid X = x\right) = \int y * p(y \mid x) dy$$

- The Law of Total Expectation or Law of Iterated Expectation:

$$E(Y) = E\left[E(Y \mid X)\right] = \int E(Y \mid X = x) p_X(x) dx$$

Some more properties

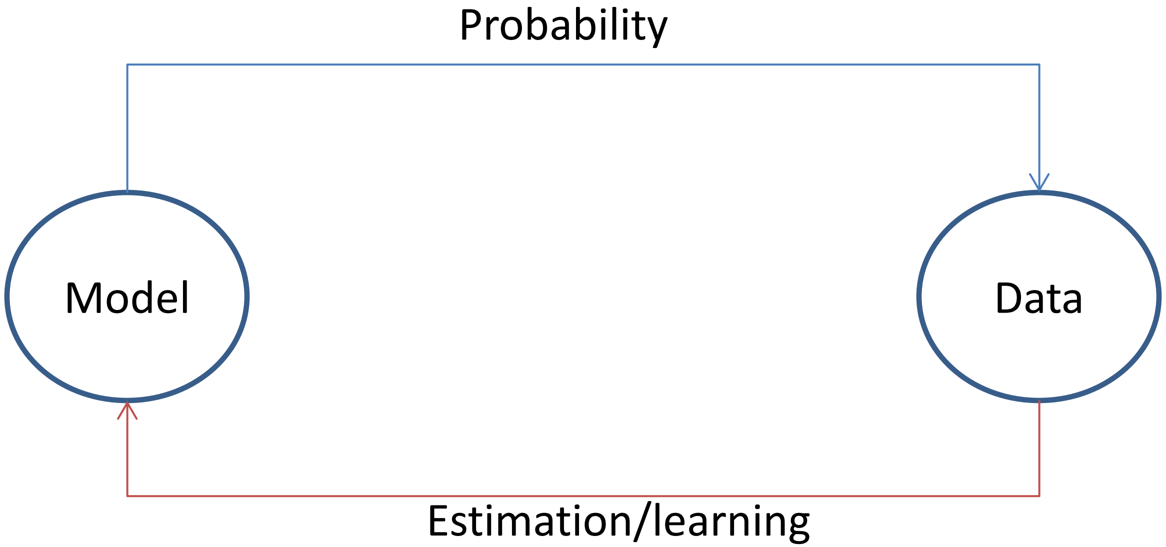
- The law of Total Variance:

$$Var(Y) = Var\left[E(Y \mid X)\right] + E\left[Var(Y \mid X)\right]$$

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Mean and Variance
- The big picture: probabilistic inference
- Examples

The Big Picture

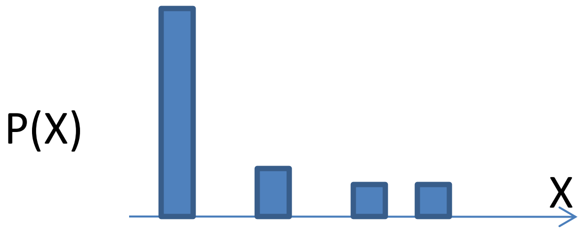


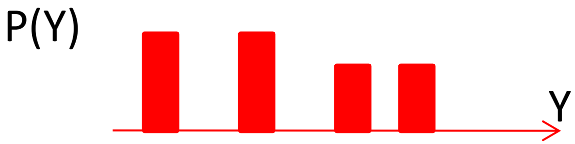
Statistical Inference

- Given observations from a model
  - What (conditional) independence assumptions hold?
    - **Structure learning**
  - If you know the family of the model (ex, multinomial), what are the value of the parameters: MLE, MAP, Bayesian estimation.
    - **Parameter learning**
      - Knowing prior or not: MAP vs. MLE

Information Theory

- P(X) encodes our uncertainty about X
  - Some variables are more uncertain that others (which?)





- How can we quantify this intuition?
  - Entropy: average number of bits required to encode X

$$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)} = -\sum_x P(x) \log P(x)$$

Information Theory cont.

- Entropy: average number of bits required to encode X

$$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)} = -\sum_x P(x) \log P(x)$$

- We can define conditional entropy similarly

$$H_P(X \mid Y) = E\left[\log \frac{1}{p(x \mid y)}\right] = H_P(X,Y) - H_P(Y)$$

  - **Interpretation**: once Y is known, we only need H(X,Y) – H(Y) bits

- We can also define chain rule for entropies (not surprising)

$$H_P(X,Y,Z) = H_P(X) + H_P(Y \mid X) + H_P(Z \mid X,Y)$$

Mutual Information: MI

- Remember independence?
  - If X⊥Y then knowing Y won’t change our belief about X
  - Mutual information can help quantify this! (not the only way though)
- MI: $I_P(X;Y) = H_P(X) - H_P(X \mid Y)$
  - “The amount of uncertainty in X which is removed by knowing Y” (or, the amount of information contained <u>mutually</u>)
  - Symmetric
  - I(X;Y) = 0 iff X and Y are independent!

$$I(X;Y) = \sum_y \sum_x p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Chi Square Test for Independence (Example)

	Republican	Democrat	Independent	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

- State the **hypotheses**

H_0: Gender and voting preferences are independent.
H_a: Gender and voting preferences are not independent

- Choose **significance level**

Say, 0.05

Chi Square Test for Independence

	Republican	Democrat	Independent	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

- Analyze sample data
  - **Degrees of freedom =**

    (|g|-1) * (|v|-1) = (2-1) * (3-1) = 2
  - **Expected frequency count =**

    $E_{g,v} = (n_g * n_v) / n$

    $E_{m,r} = (400 * 450) / 1000 = 180000/1000 = 180$
    $E_{m,d}= (400 * 450) / 1000 = 180000/1000 = 180$
    $E_{m,i} = (400 * 100) / 1000 = 40000/1000 = 40$
    $E_{f,r} = (600 * 450) / 1000 = 270000/1000 = 270$
    $E_{f,d} = (600 * 450) / 1000 = 270000/1000 = 270$
    $E_{f,i} = (600 * 100) / 1000 = 60000/1000 = 60$

Chi Square Test for Independence

	Republican	Democrat	Independent	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

- Chi-square **test statistic**

$$X^2 = \left[\sum \frac{(O_{g,v} - E_{g,v})^2}{E_{g,v}} \right]$$

- $X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40 + (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/40$
- $X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$
- $X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$

Chi Square Test for Independence

- **P-value**
  - Probability of observing a sample statistic as extreme as the test statistic
  - $P(X^2 \geq 16.2) = 0.0003$
- Since **P-value** (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis
- There is a relationship between gender and voting preference

What you must know

- ~~See the last Table in the distributed material~~
  - Probability models
  - Random variables
  - Conditional probability
  - Probability density function
  - Marginalization
  - Expectation
  - Normal distribution
  - Probabilistic inference: MLE, MAP

What you **MUST** know

- Probabilistic inference: MLE, MAP
- MLE = arg max P (X|Y)

  where X is the feature and Y is the class

  P(X|Y) is the likelihood
- MAP= arg max P (Y|X)

  P(Y|X) is posterior (*a posteriori*)

  where X is the feature and Y is the class

Acknowledgments

- Carlos Guestrin recitation slides: http://www.cs.cmu.edu/~guestrin/Class/10708/recitations/r1/Probability_and_Statistics_Review.ppt
- Andrew Moore Tutorial: http://www.autonlab.org/tutorials/prob.html
- Monty hall problem: http://en.wikipedia.org/wiki/Monty_Hall_problem
- http://www.cs.cmu.edu/~guestrin/Class/10701-F07/recitation_schedule.html
- Chi-square test for independence

  http://stattrek.com/chi-square-test/independence.aspx
- David Forsyth, unpublished review chapter (available in blackboard)