CS60050
Machine Learning
Probability Overview

Somak Aditya Sudeshna Sarkar
Department of CSE, IIT Kharagpur
Sep 8, 2023

Probability Overview

- Events
  - discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs
- Joint probability distribution
- Expectations
- Independence, Conditional independence

Probability

- Intuition:
  - In a process, several outcomes are possible
  - When the process is repeated a large number of times, each outcome occurs with a ***relative frequency***, or ***probability***
- Probability arises in two contexts
  - In actual repeated experiments
    - Example: You record the color of 1,000 cars driving by. 57 of them are green. You estimate the probability of a car being green as 57/1,000 = 0.057.
  - In idealized conceptions of a repeated process
    - Example: You consider the behavior of an unbiased six-sided die. The expected probability of rolling a 5 is 1/6 = 0.1667.
    - Example: You need a model for how people's heights are distributed. You choose a normal distribution to represent the expected relative probabilities.

- Why does Uncertainty arise?
    1. Inherent stochasticity in the system being modeled
    2. Incomplete observability
        - Even deterministic systems can appear stochastic when we cannot observe all of the variables that drive the behavior of the system
    3. Incomplete modeling
        - When we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model's predictions
        - E.g., discretization of real-numbered values, dimensionality reduction, etc.

    - Noisy measurements
    - Limited Model Complexity

Random variables

- A ***random variable*** A is a variable that can take on different values
  - Examples: X = rolling a die
    - Possible values of X comprise the **sample space**, or **outcome space**, $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
    - We denote the event of "seeing a 5" as $\{X = 5\}$ or $X = 5$
    - The probability of the event is $P(\{X = 5\})$ or $P(X = 5)$
    - Also, $P(5)$ can be used to denote the probability that X takes the value of 5
  - A = True if a randomly drawn person from our class is female
  - A = True if two randomly drawn persons from our class have same birthday
- A ***probability distribution*** is a description of how likely a random variable is to take on each of its possible states
  - A compact notation is common, where $P(X)$ is the probability distribution over the random variable X
    - Also, the notation $X \sim P(X)$ can be used to denote that the random variable X has probability distribution $P(X)$

Sample Space, Event, Random variable

- The sample space S is the set of possible outcomes of an experiment.
- Points ω in S are called sample outcomes.
- Subsets of S are called Events.
- Example. If we toss a coin twice then Ω = {HH,HT, TH, TT}.

The event that the first toss is heads is A = {HH,HT}

Axioms of probability

- The probability of an event A in the given sample space $\mathcal{S}$, denoted as $P(A)$, must satisfies the following properties:
  1. **Non-negativity**
     - For any event $\mathrm{A} \in \mathcal{S}$, $P(A) \geq 0$
  2. **All possible outcomes**
     - Probability of the entire sample space is 1, $P(\mathcal{S}) = 1$
  3. **Additivity of disjoint events**
     - For all events $\mathrm{A}_1, A_2 \in \mathcal{S}$ that are mutually exclusive ($\mathrm{A}_1 \cap A_2 = \emptyset$), the probability that both events happen is equal to the sum of their individual probabilities, $P(\mathrm{A}_1 \vee \mathrm{A}_2) = P(A_1) + P(\mathrm{A}_2)$
- The probability of a random variable $P(X)$ must obey the axioms of probability over the possible values in the sample space $\mathcal{S}$

Random Variable

- A random variable is a function defined over the sample space. a random variable (r.v.) X denotes possible outcomes of an event.
- Can be discrete (i.e., finite many possible outcomes) or continuous
  - Some examples of discrete r.v.

    $X \in \{0, 1\}$ denoting outcomes of a coin-toss

    $X \in \{1, 2, \ldots, 6\}$ denoting outcome of a dice roll
  - Some examples of continuous r.v.

    $X \in (0, 1)$ denoting the bias of a coin

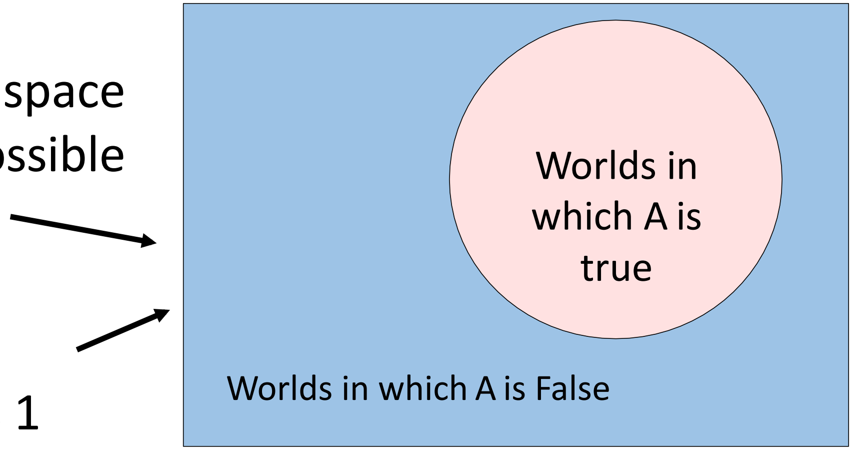
    $X \in \mathbb{R}$ denoting heights of students in CS771

    $X \in \mathbb{R}$ denoting time to get to your hall from the department

Visualizing A

Sample space of all possible worlds

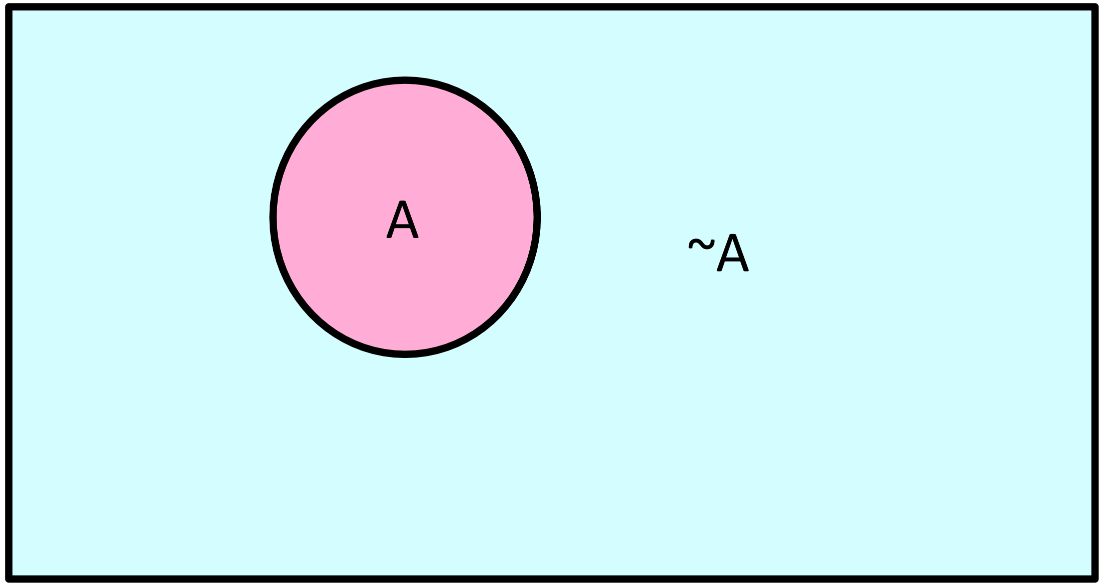
Its area is 1



P(A) = Area of reddish oval

Elementary Probability in Pictures

P(∼A) + P(A) = 1



A useful theorem

0 <= P(A) <= 1, P(True) = 1, P(False) = 0,

P(AV B) = P(A) + P(B) – P(A ∧ B)

$$P(A) = P(A \wedge B) + P(A \wedge \sim B)$$

$$A = [A \wedge (B \vee \sim B)] = [(A \wedge B) \vee (A \wedge \sim B)]$$

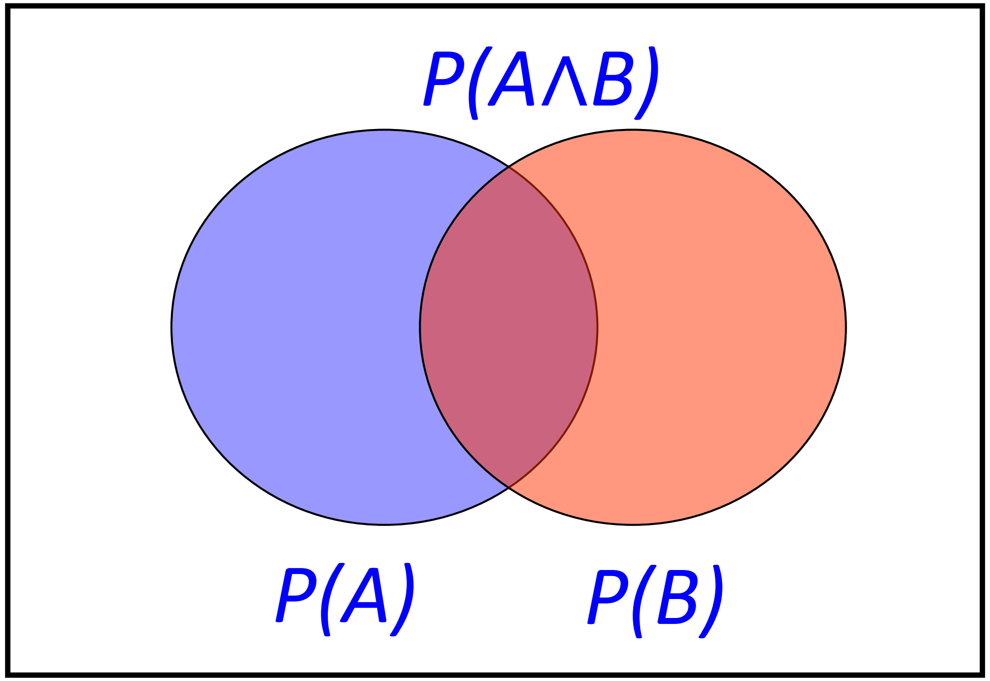
$$P(A) = P(A \wedge B) + P(A \wedge \sim B) - P\big((A \wedge B) \wedge (A \wedge \sim B)\big)$$

$$= P(A \wedge B) + P(A \wedge \sim B) - \cancel{P(A \wedge B \wedge A \wedge \sim B)}$$

Elementary Probability | Conditional Probability

P(A) = P(A ∧ B) + P(A ∧ ~B)

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



Corollary: The Chain Rule

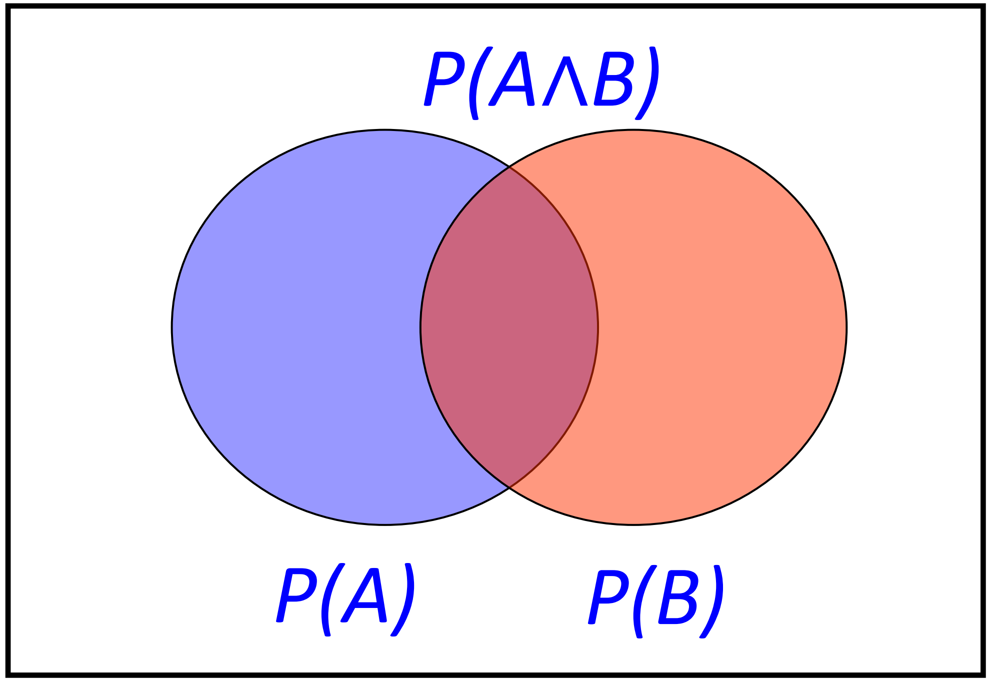
$$P(A \wedge B) = P(A|B)P(B)$$

Bayes Rule

$$P(A \wedge B) = P(B|A) * P(A) = P(A|B) * P(B)$$

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

- $P(A)$, the prior probability
- $P(A|\ B)$, the posterior probability
- $P(B|\ A)$, the likelihood of $\boldsymbol{B}$ given A
- P(B), the evidence

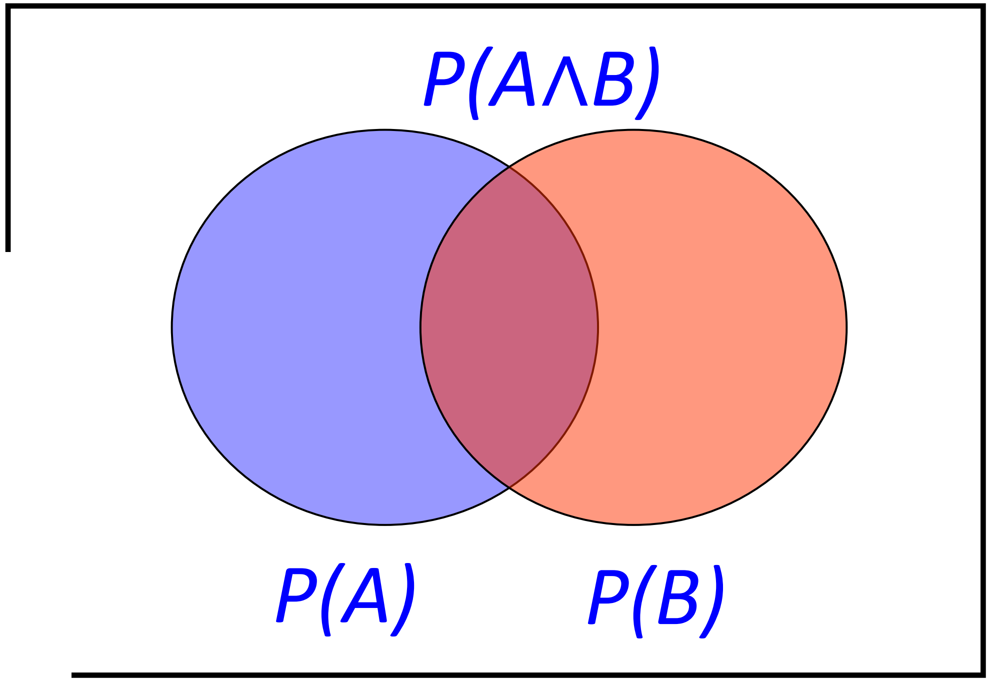


Bayes Rule

$$P(A \wedge B) = P(B|A) * P(A) = P(A|B) * P(B)$$

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

Other Forms of Bayes Rule



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

$$P(A|B \wedge X) = \frac{P(B|A \wedge X) * P(A \wedge X)}{P(B \wedge X)}$$

Applying Bayes Rule

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

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

Assume:

- $P(A) = 0.05$
- $P(B|A) = 0.8$
- $P(B|{\sim}A) = 0.2$

$$P(A|B) = \frac{0.8 \times 0.05}{0.8 \times 0.05 + 0.2 \times 0.95} {\sim} 0.17$$

- what is P(flu | cough) = P(A|B)?

Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

$$\boxed{P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}}$$

- Example:
  - M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\} \text{Example givens}$$

$$\boxed{P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}}$$

  - Note: posterior probability of meningitis still very small
  - Note: you should still get stiff necks checked out! Why?

Why we are learning this?

$$x \longrightarrow \boxed{h} \longrightarrow y$$

Hypothesis

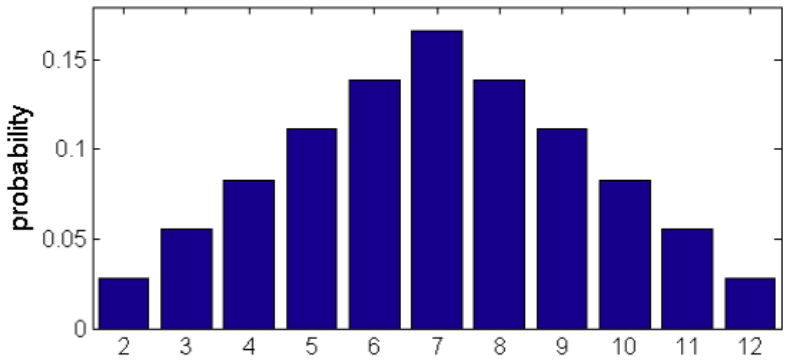
Learn $P(Y|X)$

- A *probability distribution* is a description of how likely a random variable is to take on each of its possible states
  - $P(X)$ denotes the probability distribution over the random variable X
    - The notation $\mathrm{X} \sim P(X)$ can be used to denote that the random variable X has probability distribution $P(X)$
- Random variables can be discrete or continuous
  - Discrete random variables have finite number of states: e.g., the sides of a die
  - Continuous random variables have infinite number of states: e.g., the height of a person

Probability Distribution

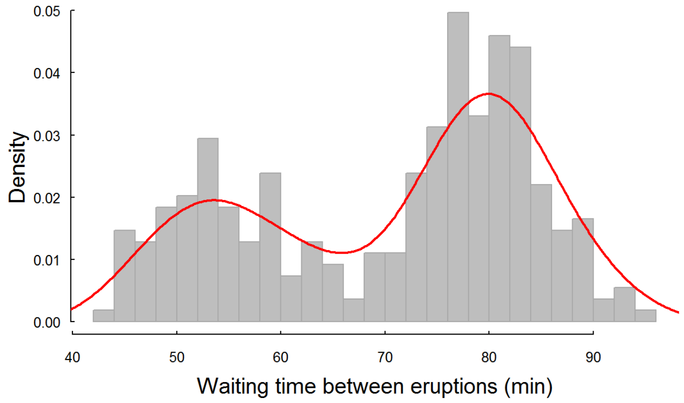
A probability distribution over discrete variables may be described using a *probability mass function* (PMF)

- E.g., sum of two dice



- A probability distribution over continuous variables m be described using a *probability density function* (PDF)
  - E.g., waiting time between eruptions of Old Faithful
  - A PDF gives the probability of an infinitesimal region with volume δX
  - To find the probability over an interval [*a*, *b*], we can integrate the PDF as follows:

$$P(X \in [a,b]) = \int_a^b P(X)dX$$



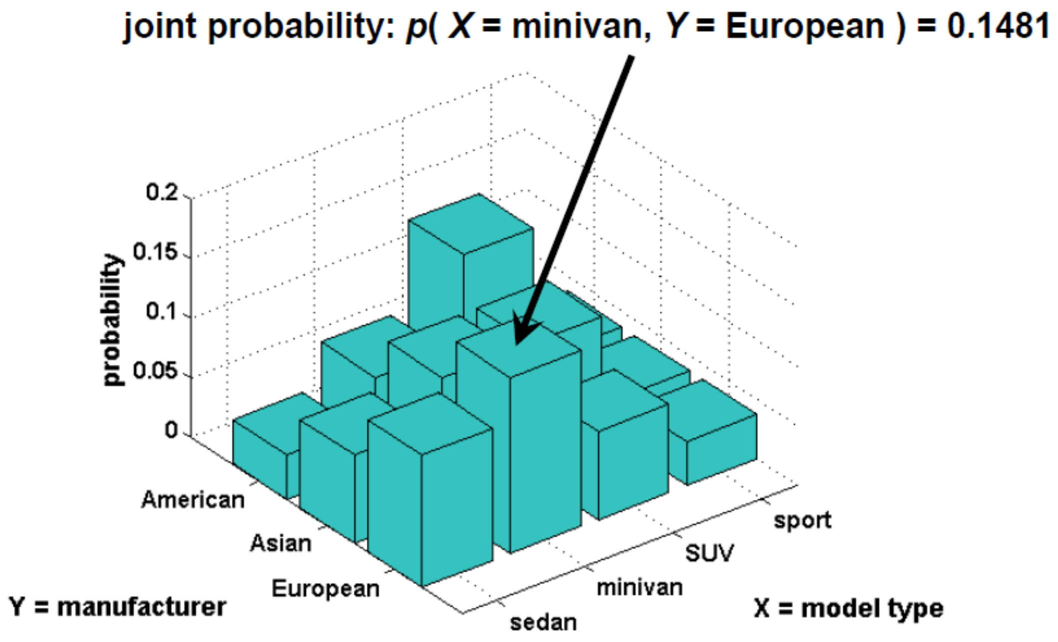
Multivariate Random Variables

- Probability distributions defined over multiple random variables
  - joint, conditional, and marginal probability distributions

- A ***multivariate random variable*** is a vector of multiple random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$

Joint Probability Distribution

- Probability distribution over multiple variables is known as a ***joint probability distribution***
- Given any values *x* and *y* of two random variables X and Y, what is the probability that X = *x* and Y = *y* simultaneously?
  - $P(X = x, Y = y)$ denotes the joint probability
  - We may also write $P(x, y)$



$$\sum_x \sum_y p(X = x, Y = y) = 1$$

Independent Events

- Definition: two events A and B are *independent* if
$$P(A \wedge B) = P(A) * P(B)$$
- Intuition: knowing A tells us nothing about the value of B (and vice versa)

Marginal Probability Distribution

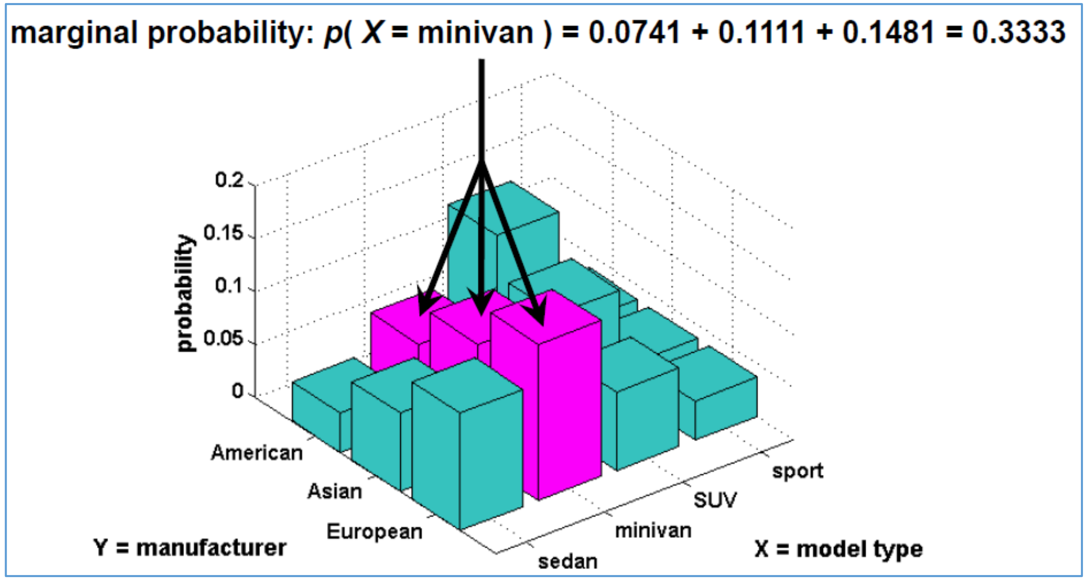
Marginal probability distribution is the probability distribution of a single variable

It is calculated based on the joint probability distribution $P(X, Y)$ using the sum rule:

$$P(X = x) = \sum_y P(X = x, Y = y)$$

- For continuous random variables, the summation is replaced with integration,

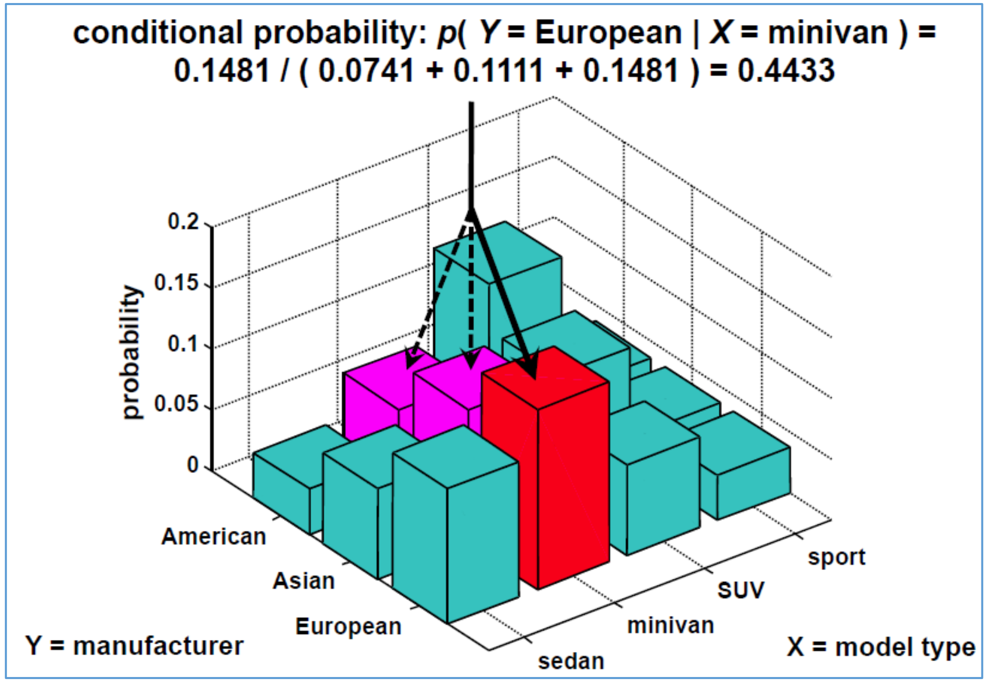
$$P(X = x) = \int P(X = x, Y = y)\, dy$$



Conditional Probability Distribution

Conditional probability distribution is the probability distribution of one variable provided that another variable has taken a certain value.

$$P(X = x | Y = y) = \frac{P(X = x,\ \ Y = y)}{P(Y = y)}$$



Expected Value

- The ***expected value*** or ***expectation*** of a function $f(X)$ with respect to a probability distribution $P(X)$ is the average (mean) when X is drawn from $P(X)$
- For a discrete random variable *X*, it is calculated as

$$\mathbb{E}_{X \sim P}[f(X)] = \sum_{X} P(X) f(X)$$

- For a continuous random variable *X*, it is calculated as

$$\mathbb{E}_{X \sim P}[f(X)] = \int P(X) f(X) \, dX$$

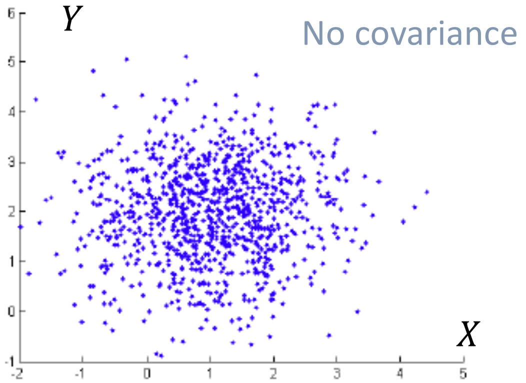
- When the identity of the distribution is clear from the context, we can write $\mathbb{E}_X[f(X)]$
- If it is clear which random variable is used, we can write just $\mathbb{E}[f(X)]$

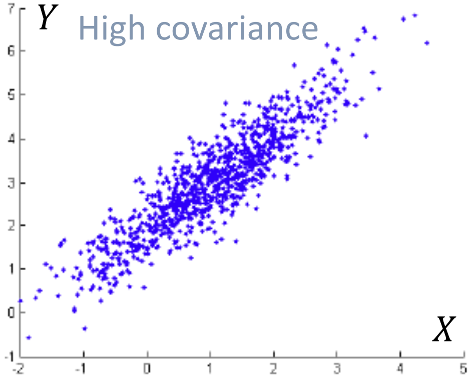
Covariance

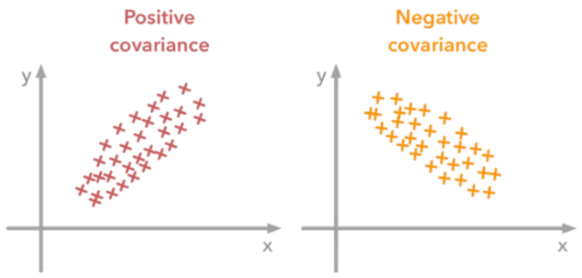
- *Covariance* gives the measure of how much two random variables are linearly related to each other

$$\mathrm{Cov}\big(f(X), g(Y)\big) = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])(g(Y) - \mathbb{E}[g(Y)])]$$

- The covariance measures the tendency for X and Y to deviate from their means in same (or opposite) directions at same time







Picture from: Jeff Howbert — Machine Learning Math Essentials

Covariance Matrix

Covariance matrix of a multivariate rv $\mathbf{X}$ with states $\mathbf{x} \in \mathbb{R}^{\boldsymbol{n}}$ is an $n \times n$ matrix, such that

$$\mathrm{Cov}(\mathbf{X})_{i,j} = \mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j)$$

I.e.,

$$\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_{\mathbf{1}}) & \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_{\mathbf{2}}) & \cdots & \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_{\boldsymbol{n}}) \\ \mathrm{Cov}(\mathbf{x}_2, \mathbf{x}_{\mathbf{1}}) & \ddots & & \mathrm{Cov}(\mathbf{x}_2, \mathbf{x}_{\boldsymbol{n}}) \\ \vdots & & & \vdots \\ \mathrm{Cov}(\mathbf{x}_n, \mathbf{x}_{\mathbf{1}}) & \mathrm{Cov}(\mathbf{x}_n, \mathbf{x}_{\mathbf{2}}) & \cdots & \mathrm{Cov}(\mathbf{x}_{\boldsymbol{n}}, \mathbf{x}_{\boldsymbol{n}}) \end{bmatrix}$$

The diagonal elements of the covariance matrix are the variances of the vector elements

$$\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_i) = \mathrm{Var}(\mathbf{x}_i)$$

- Note that the covariance matrix is symmetric, since $\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{Cov}(\mathbf{x}_j, \mathbf{x}_i)$

Some Basic Rules

Sum Rule: Gives the marginal probability distribution from joint probability distribution

For discrete r.v.: $p(X) = \sum_Y p(X, Y)$

For continuous r.v.: $p(X) = \int_Y p(X, Y)dY$

- **Product Rule:** $p(X, Y) = p(Y|X)p(X) = p(X|Y)p(Y)$
- **Bayes' rule:** Gives conditional probability distribution (can derive it from product rule)

$$\boxed{p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}}$$

For discrete r.v.: $p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$

For continuous r.v.: $p(Y|X) = \frac{p(X|Y)p(Y)}{\int_Y p(X|Y)p(Y)dY}$

- **Chain Rule:** $p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2|X_1)\ldots p(X_N|X_1, \ldots, X_{N-1})$

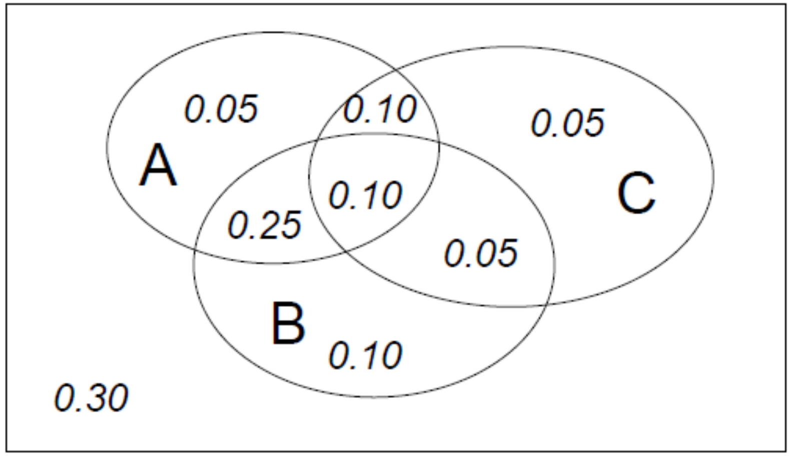
The Joint Distribution

Example: Boolean variables A, B, C

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

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations. (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. Probability must sum to 1.



Using joint distribution

Can ask for any logical expression involving these variables

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E) = \sum_{\text{rows matching E}} P(\text{row})$$

- $P(E_1|E_2) = \frac{\sum_{\text{rows matching } \mathrm{E}_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$

Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E)=\sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E)=\sum_{\text{rows matching } E} P(\text{row})$$

Using the joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E)=\sum_{\text{rows matching } E} P(\text{row})$$

Inference

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Inference

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Learning and the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Suppose we want to learn the function f: <G, H> $\rightarrow$W

Equivalently, P(W | G, H)

Solution: learn joint distribution from data, calculate P(W | G, H) e.g.,

P(W=rich | G = female, H = 40.5-) =

[A. Moore]

The solution to learn $P(Y|X)$?

- Main problem: learning $P(Y|X)$ may require more data than we have
- Say, learning a joint distribution with 100 attributes

- # of rows in this table?

$$2^{100} \geq 10^{30}$$

- # of people on earth?

$$10^{9}$$

- fraction of rows with 0 training examples?

Slide credit: Tom Mitchell

What to do?

1. Be smart about **how we estimate probabilities from sparse data**
   – maximum likelihood estimates
   – maximum a posteriori estimates

2. Be smart about **how to represent joint distributions**
   – Bayes networks, graphical models

1. Be smart about how we estimate probabilities

Estimating the probability

- Flip the coin repeatedly, observing
  - It turns heads α_1 times
  - It turns tails α_0 times
- Your estimate for $P(X = 1)$ is?

- Case A: 100 flips: 51 Heads ($X = 1$), 49 Tails ($X = 0$)

  $P(X = 1) = ?$
- Case B: 3 flips: 2 Heads ($X = 1$), 1 Tails ($X = 0$)

  $P(X = 1) = ?$
- Case C: (online learning) keep flipping, want single learning algorithm that gives reasonable estimate after each flip

Slide credit: Tom Mitchell

Two principles for estimating parameters

- **Maximum Likelihood Estimate (MLE)**
  Choose θ that maximizes probability of observed data

$$\hat{\boldsymbol{\theta}}^{\mathrm{MLE}} = \underset{\theta}{\operatorname{argmax}}\, P(Data|\theta)$$

- **Maximum a posteriori estimation (MAP)**
  Choose θ that is most probable given prior probability and data

$$\hat{\boldsymbol{\theta}}^{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}}\, P(\theta|D) = \underset{\theta}{\operatorname{argmax}} \frac{P(Data|\theta)P(\theta)}{P(Data)}$$

Slide credit: Tom Mitchell

Two principles for estimating parameters

- **Maximum Likelihood Estimate (MLE)**
  Choose θ that maximizes $P(Data|\theta)$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

- **Maximum a posteriori estimation (MAP)**
  Choose θ that maximize $P(\theta|Data)$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{MAP}} = \frac{(\alpha_1 + \#\text{hallucinated 1s})}{(\alpha_1 + \#\text{hallucinated 1}s) + (\alpha_0 + \#\text{hallucinated 0s})}$$

Slide credit: Tom Mitchell

Maximum likelihood estimate

$X = 1 \quad X = 0$

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

- Each flip yields Boolean value for X

$$X \sim \text{Bernoulli}: P(X) = \theta^X (1 - \theta)^{1-X}$$

- Data set D of independent, identically distributed (iid) flips, produces α_1 ones, α_0 zeros

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

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\theta} P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Slide credit: Tom Mitchell

Maximum Likelihood Estimate for Θ

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

$$\theta^* = \arg\max_\theta P(D|\theta) = \arg\max_\theta \ln P(D|\theta)$$

$$= \arg\max_\theta \ln \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

Set derivative to zero

$$\frac{\partial}{\partial\theta} \ln P(D|\theta) = 0$$

[C. Guestrin]

Example (Bernoulli trials). If the experiment consists of n Bernoulli trial with success probability θ, then

$$L(\theta|x) = \theta^{x_1}(1-\theta)^{1-x_1} \ldots \theta^{x_n}(1-\theta)^{1-x_n}$$
$$= \theta^{(x_1+\cdots+x_n)}\,(1-\theta)^{n-(x_1+\cdots+x_n)}$$

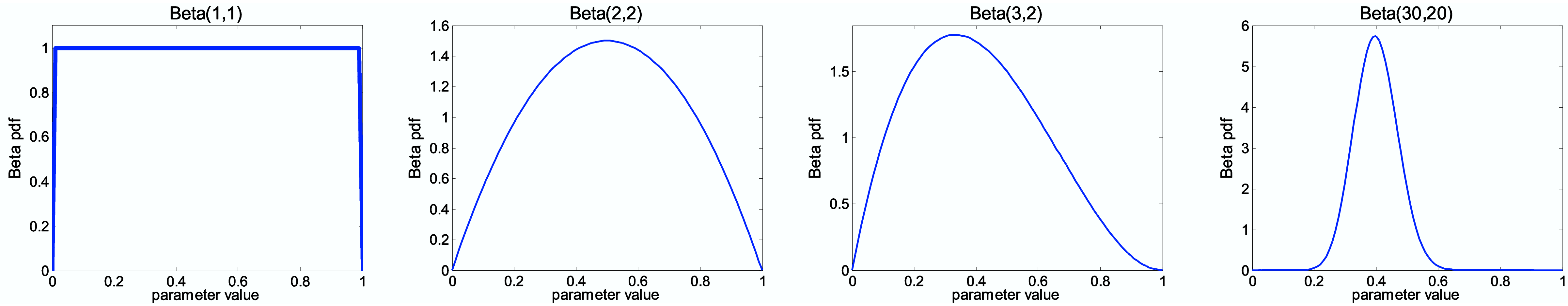
$$\ln \mathbf{L}(\theta|\mathbf{x}) = \ln\theta\left(\sum_{i=1}^{n} x_i\right) + \ln(1-\theta)\left(n - \sum_{i=1}^{n} x_i\right)$$
$$= n\bar{x}\ln\theta + n(1-\bar{x})\ln(1-\theta)$$

$$\frac{\partial}{\partial\theta}\ln \mathbf{L}(\theta|\mathbf{x}) = n\left(\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta}\right)$$

This equals zero when $\theta = \bar{x}$.

Beta prior distribution $P(\theta)$

$$P(\theta) = Beta(\beta_1, \beta_0) = \frac{1}{B(\beta_1, \beta_0)} \theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}$$



Slide credit: Tom Mitchell

Maximum likelihood estimate

$X = 1$ $X = 0$

- Data set D of iid flips, produces α_1 ones, α_0 zeros

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

- Assume prior (Conjugate prior: Closed form representation of posterior)

$$P(\theta) = Beta(\beta_1, \beta_0) = \frac{1}{B(\beta_1, \beta_0)}\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}$$

$$\widehat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)\,\mathrm{P}(\theta) = \frac{\alpha_1 + \beta_1 - 1}{(\alpha_1 + \beta_1 - 1) + (\alpha_0 + \beta_0 - 1)}$$

Slide credit: Tom Mitchell

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

- choose parameters θ that maximize $P(\text{data} \mid \theta)$

Principle 2 (maximum a posteriori prob.):

- choose parameters θ that maximize

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

Some terminology

- Likelihood function: P(data | θ)
- Prior: P(θ)
- Posterior: P(θ | data)

- Conjugate prior: P(θ) is the conjugate prior for likelihood function P(data |θ) if the forms of P(θ) and P(θ | data) are the same.

You should know

- Probability basics
  - random variables, conditional probs, …
  - Bayes rule
  - Joint probability distributions
  - calculating probabilities from the joint distribution
- Estimating parameters from data
  - maximum likelihood estimates
  - maximum a posteriori estimates
  - distributions – binomial, Beta, Dirichlet, …
  - conjugate priors