A Bayesian View of ML

Bayes Rule beyond Classification

Today's Agenda

AIML

Optimal Classifier

Bayes Rule for Classification

Optimal Decisions

Bayes Rule for Decision Making

Best Estimate of the World

Bayes Rule for Learning

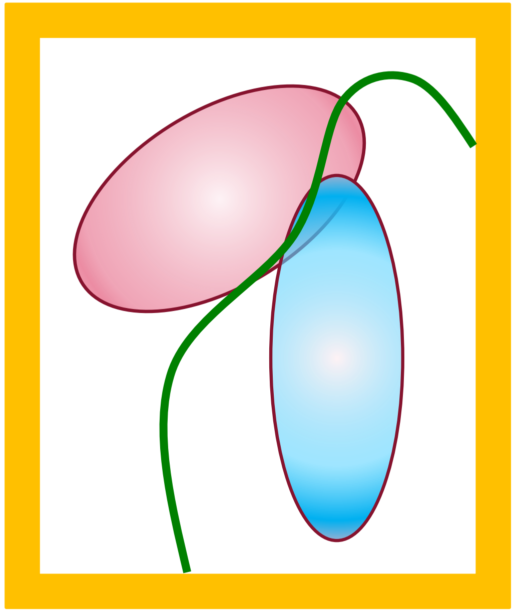
Naïve Bayes Classifier

Uncorrelated World

Understanding Causation

Probabilistic Graphical Models

Partially correlated World

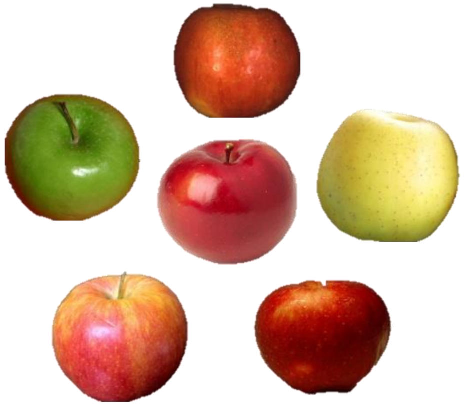


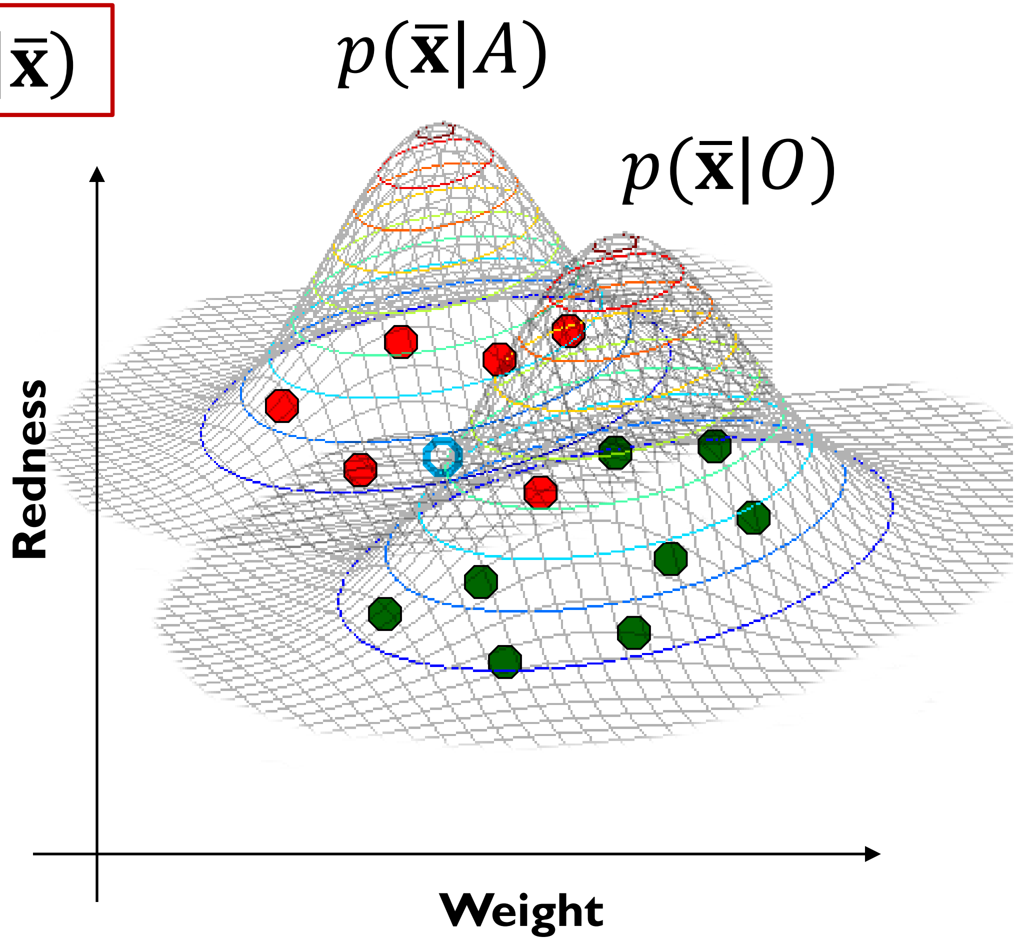
Bayes Classifier

A RECAP

Feature Space: Bayes Classifier

$\boxed{P(A|\bar{\mathbf{x}}) \;\; vs \; P(O|\bar{\mathbf{x}})}$





Feature Space Representation

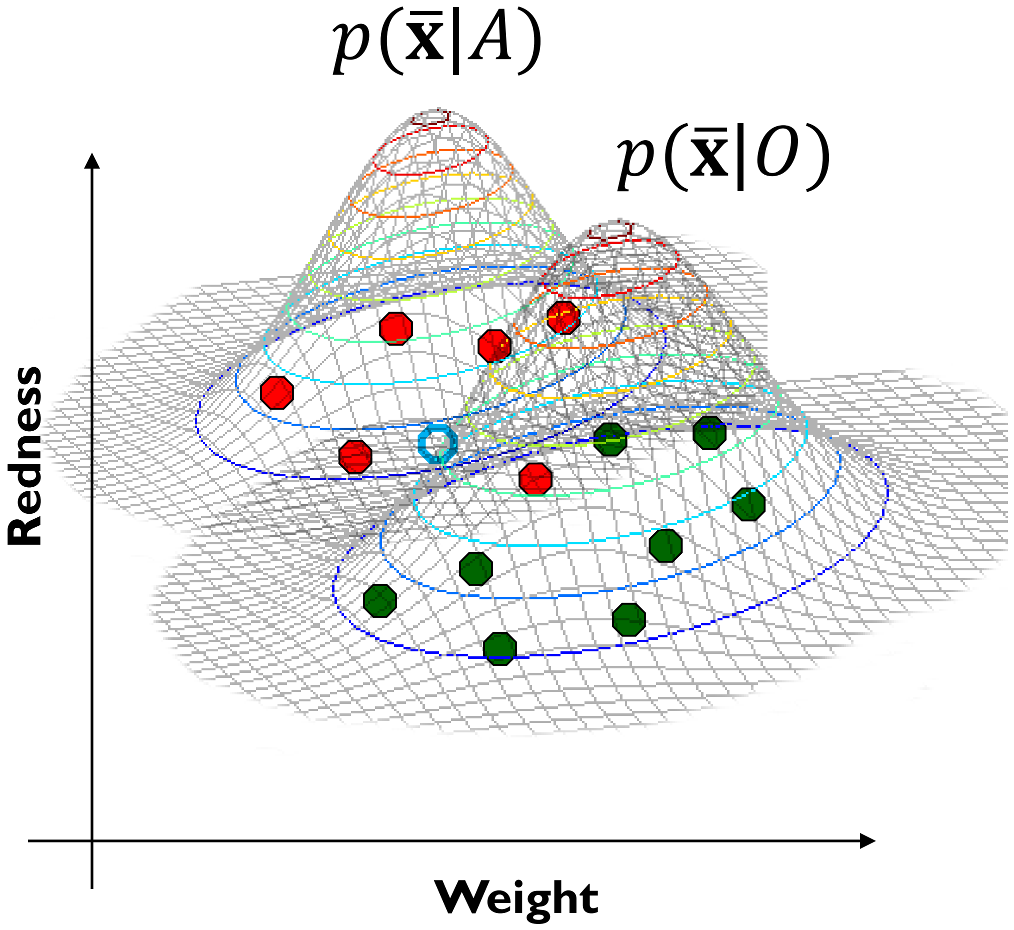
Feature Space: Bayes Classifier

- Compute the likelihoods:
- Compute the posteriors:

$$P(A|\bar{\mathbf{x}}) \;\; vs \; P(O|\bar{\mathbf{x}})$$

$$\frac{p(\bar{\mathbf{x}}|A) \times P(A)}{p(\bar{\mathbf{x}})} \;\; vs \;\; \frac{p(\bar{\mathbf{x}}|O) \times P(O)}{p(\bar{\mathbf{x}})}$$

- Assign class with highest posterior probability



Feature Space Representation

Summary/Comments on Bayes Rule

- We should consider the prior probability (belief) along with the likelihood (observation) to arrive at the optimal decision

$$P(A|\bar{\mathbf{x}}) = \frac{p(\bar{\mathbf{x}}|A) \times P(A)}{p(\bar{\mathbf{x}})}$$

- One can ignore the data evidence in decision making
- Assuming equal prior results in Maximum Likelihood Classification

Example:

- Likelihood of **rain** given that we **hear water drops** fall (is this classification?)

Reverend Thomas Bayes, F.R.S. (1701–1761)

Topics Outline

AIML

Optimal Classifier

Bayes Rule for Classification

Optimal Decisions

Bayes Rule for Decision Making

Best Estimate of the World

Bayes Rule for Learning

Naïve Bayes Classifier

Uncorrelated World

Understanding Causation

Probabilistic Graphical Models

Partially correlated World

Bayesian Decision Making

AIML

- The rule is applicable to other decision making problems as well
- The most likely *hypothesis* after observing any data is the one that maximizes the product of *prior belief* and *likelihood of observing the data* under the hypothesis
- If we know the cost of taking an *action* under a hypothesis, $\lambda(\alpha_i|h_j)$, we can combine it to take the least cost (best) action
- We can also compute the overall Risk

$$P(h|d) = \frac{p(d|h) \times P(h)}{p(d)}$$

h : hypothesis, d : data

$$R(\alpha_i|d) = \sum_{j=1}^{c} \lambda(\alpha_i|h_j) P(h_j|d)$$

$$R = \int R(\alpha_i|x) p(x) dx$$

Topics Outline

Optimal Classifier

Bayes Rule for Classification

Optimal Decisions

Bayes Rule for Decision Making

Best Estimate of the World

Bayes Rule for Learning

Naïve Bayes Classifier

Uncorrelated World

Understanding Causation

Probabilistic Graphical Models

Partially correlated World

Bayesian Parameter Estimation

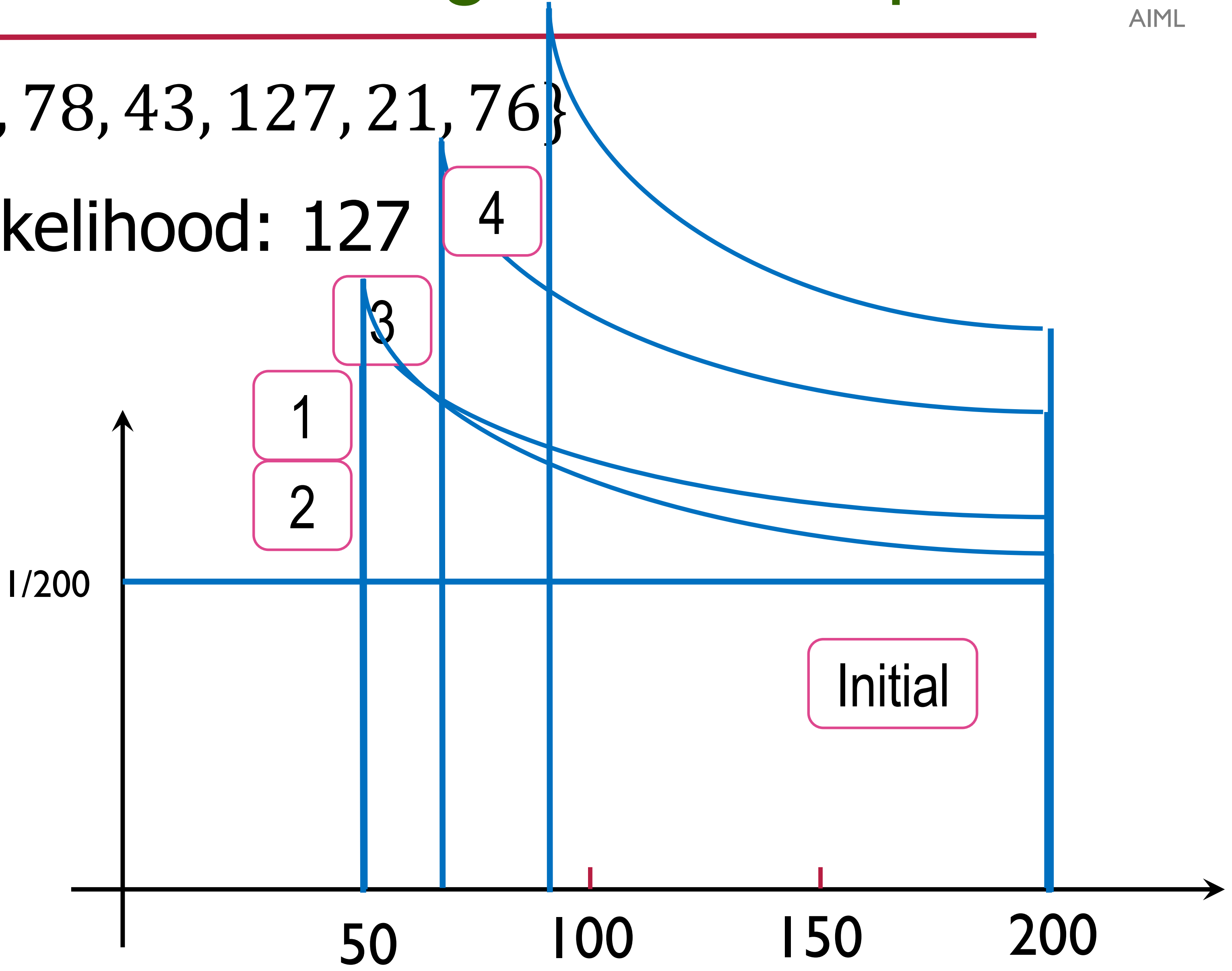
- What is the width of a canal: $U[0, X]$
- What is X?
  - Maximum width of a canal
  - Frequentist: It is an unknown constant in $[0, X]$
  - Bayesian view: It is an unknown RV $U[0, X]$
    - Modelled as a density
- How do you estimate X?
  - Measure random canal widths at random points.
    - Observations: $\{X_1, X_2, X_3, X_4 \dots X_N\}$

Recursive Bayesian Learning: An Example

Observations: $\{50, 18, 65, 78, 43, 127, 21, 76\}$

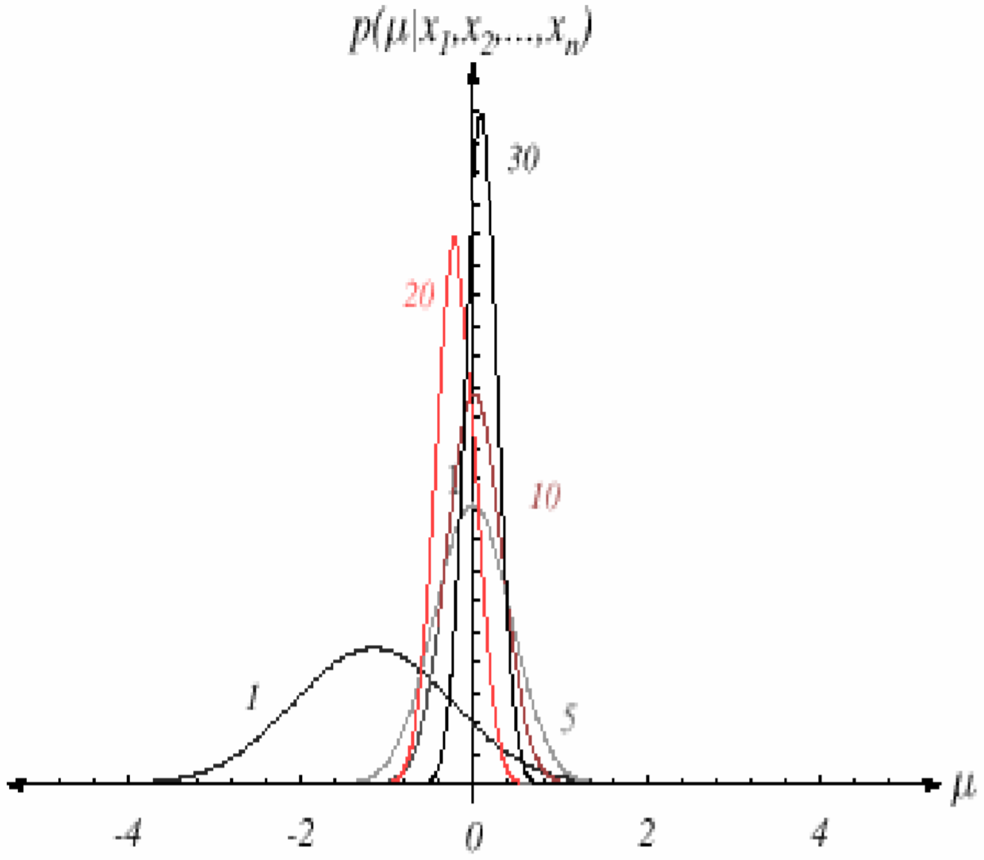
Frequentist: Maximum Likelihood: 127

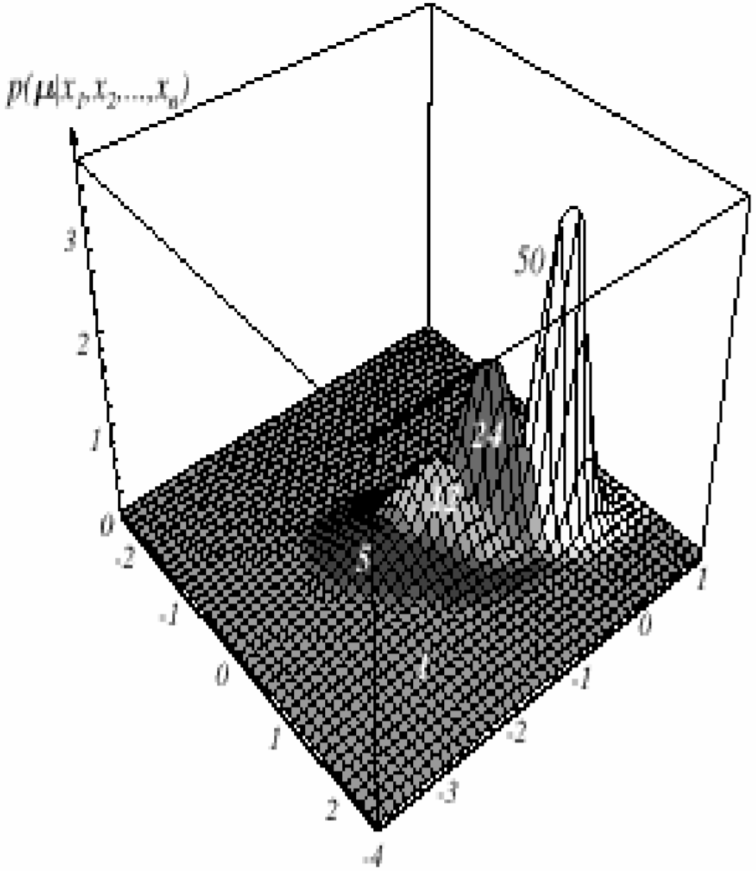
Bayesian:



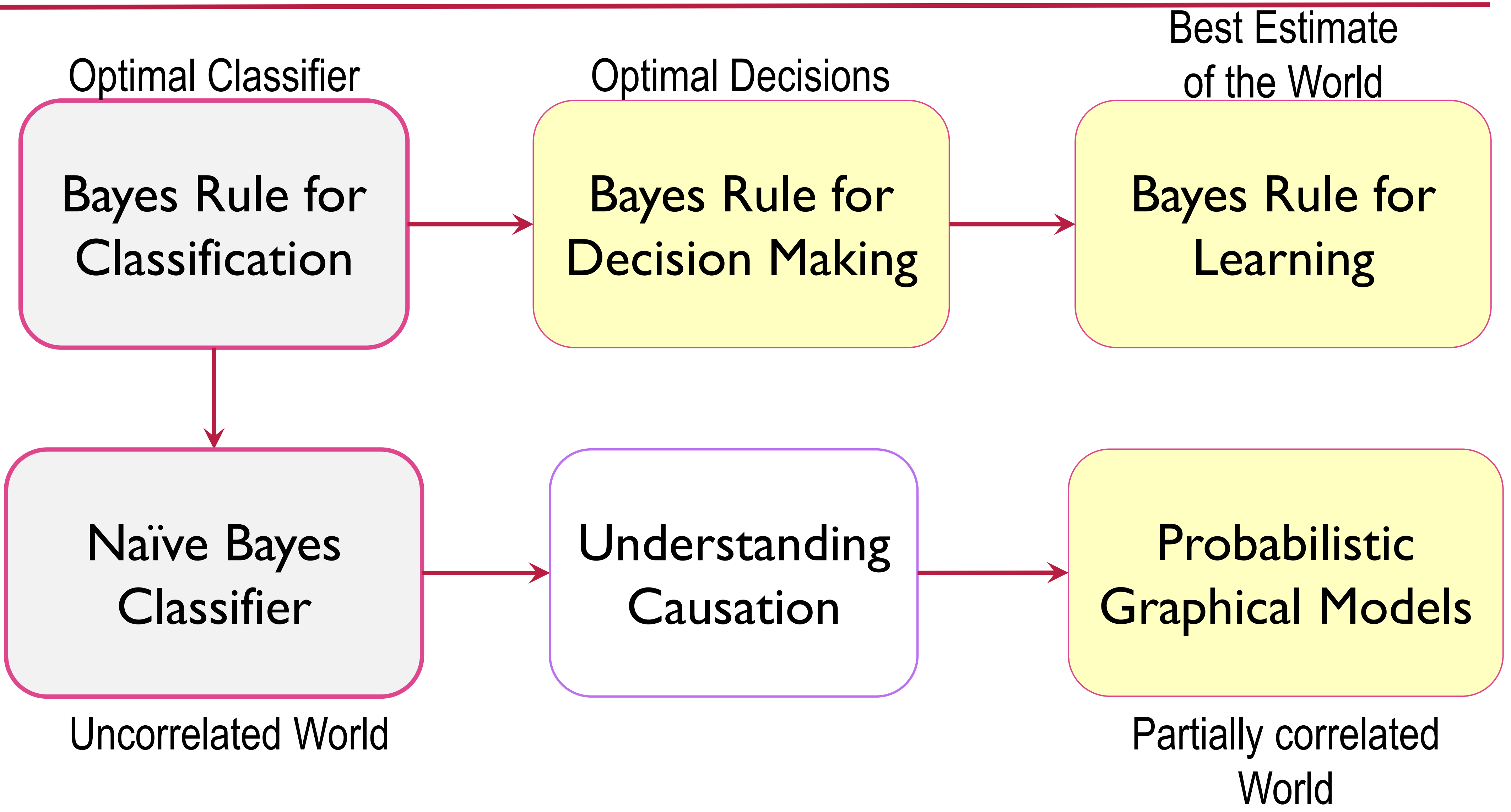
Recursive Bayesian Learning

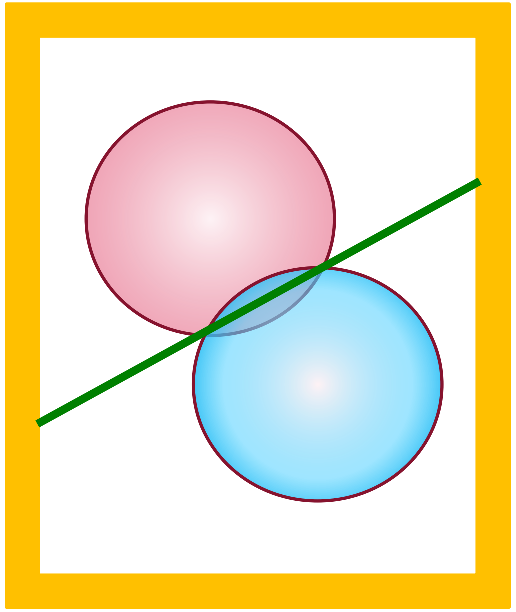
Estimating Mean of a set of variables (not necessarily Gaussian)





Topics Outline





From Bayes to Naïve Bayes

A RECAP

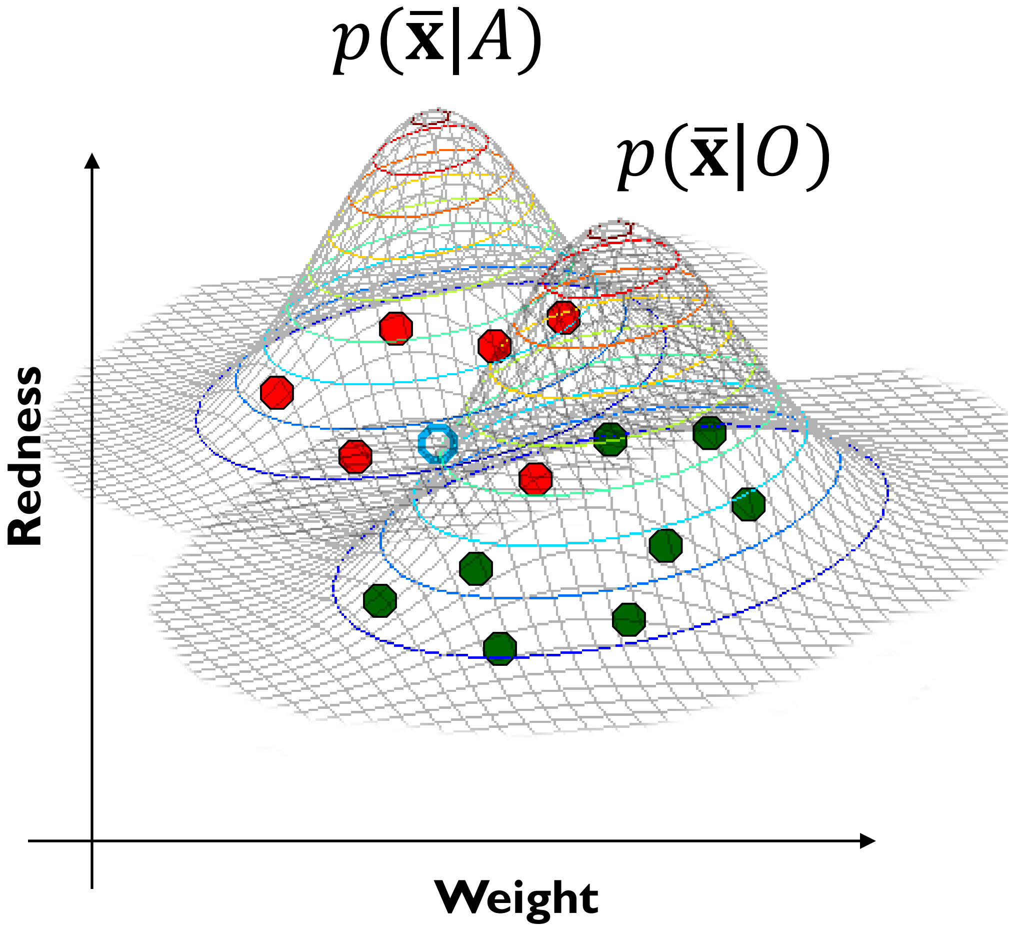
Feature Space: Bayes Classifier

- Compute the likelihoods:
- Compute the posteriors:

$$P(A|\bar{\mathbf{x}}) \;\; vs \;\; P(O|\bar{\mathbf{x}})$$

$$\frac{p(\bar{\mathbf{x}}|A) \times P(A)}{p(\bar{\mathbf{x}})} \;\; vs \;\; \frac{p(\bar{\mathbf{x}}|O) \times P(O)}{p(\bar{\mathbf{x}})}$$

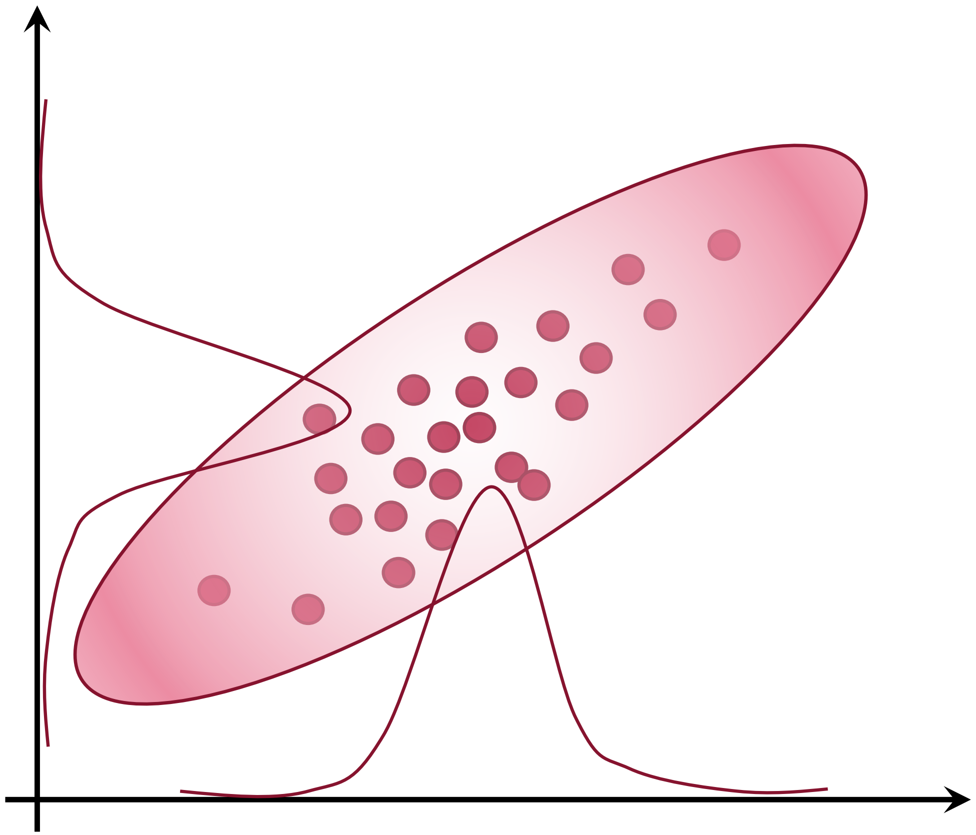
- Assign class with highest posterior probability



Feature Space Representation

What about Multiple Dimensions?

- We have variances along each dimension
- The samples also co-vary.
  i.e, features are not independent
- Captured using a covariance matrix



$$\begin{bmatrix} V_a & C_{a,b} & C_{a,c} & C_{a,d} & C_{a,e} \\ C_{a,b} & V_b & C_{b,c} & C_{b,d} & C_{b,e} \\ C_{a,c} & C_{b,c} & V_c & C_{c,d} & C_{c,e} \\ C_{a,d} & C_{b,d} & C_{c,d} & V_d & C_{d,e} \\ C_{a,e} & C_{b,e} & C_{c,e} & C_{d,e} & V_e \end{bmatrix}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$$

Likelihood Function

- $N(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

becomes

- $N(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$

Challenge of Data

$$\begin{bmatrix} V_a & C_{a,b} & C_{a,c} & C_{a,d} & C_{a,e} \\ C_{a,b} & V_b & C_{b,c} & C_{b,d} & C_{b,e} \\ C_{a,c} & C_{b,c} & V_c & C_{c,d} & C_{c,e} \\ C_{a,d} & C_{b,d} & C_{c,d} & V_d & C_{d,e} \\ C_{a,e} & C_{b,e} & C_{c,e} & C_{d,e} & V_e \end{bmatrix}$$

- 1-dim had 2 parameters to estimate
- d-dim will have not just $2d$, but over $d^2/2$ parameters.

Solving the Challenge

- Assume $\boldsymbol{\Sigma}$ to be diagonal
- i.e., features are independent
- We lose some [A LOT OF] information about the data

$$\begin{bmatrix} V_a & & & & \\ & V_b & & 0 & \\ & & V_c & & \\ & 0 & & V_d & \\ & & & & V_e \end{bmatrix}$$

Simpler likelihood

- $p(\mathbf{x}|A) = p(x_1|A) \times p(x_2|A) \times p(x_3|A) \times \cdots \times p(x_d|A)$
- A multivariate density $p(\mathbf{x}|A)$ is approximated with the product of d univariate densities: $p(x_i|A)$.
- Equivalent to assuming diagonal covariance for Normal density.
- Otherwise, Naïve Bayes classifier is same as a regular Bayesian Classifier
- Note that the univariate densities need not all be Gaussian or even the same in Naïve Bayes

Topics Outline

Optimal Classifier

Bayes Rule for Classification

Optimal Decisions

Bayes Rule for Decision Making

Best Estimate of the World

Bayes Rule for Learning

Naïve Bayes Classifier

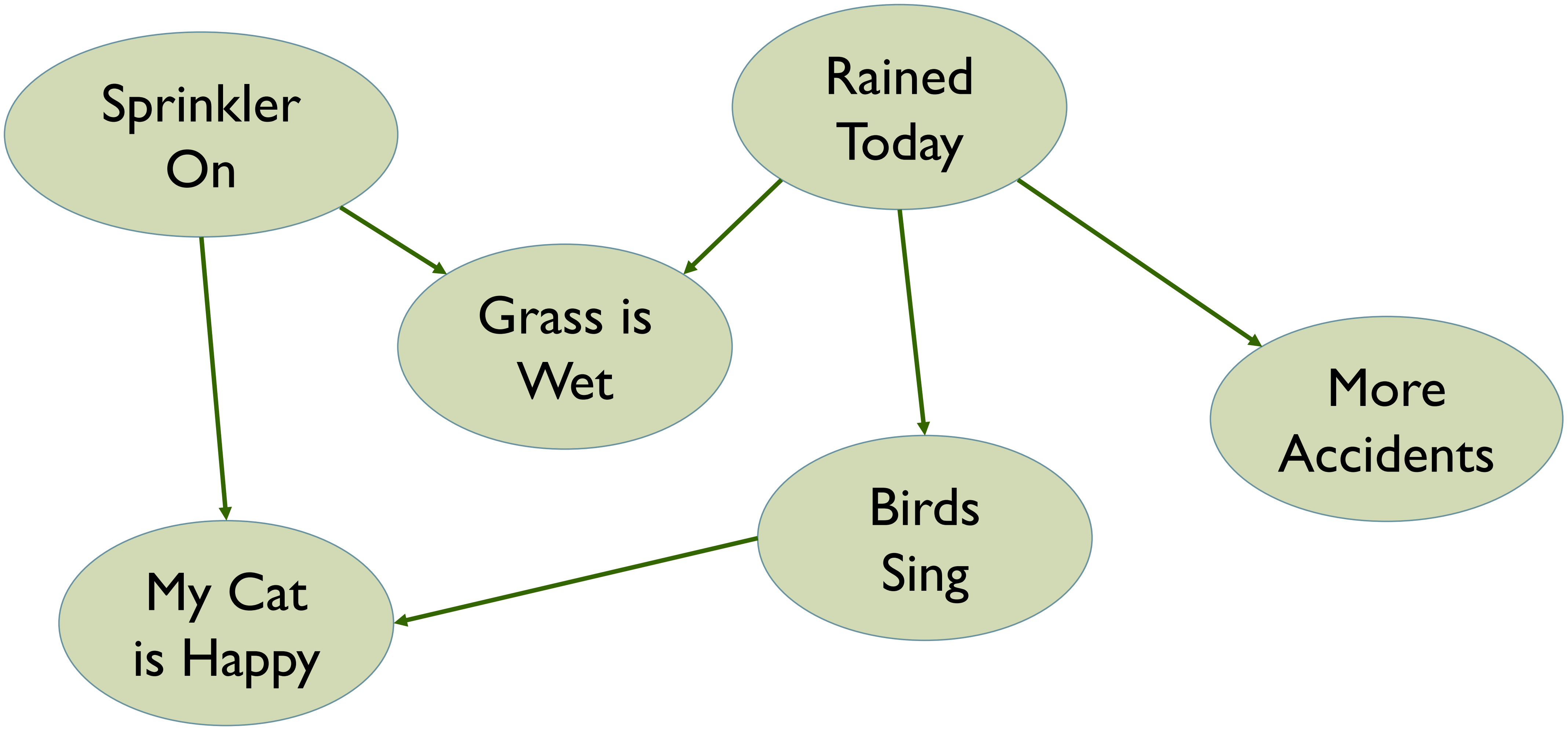
Uncorrelated World

Understanding Causation

Probabilistic Graphical Models

Partially correlated World

Causality vs Correlation

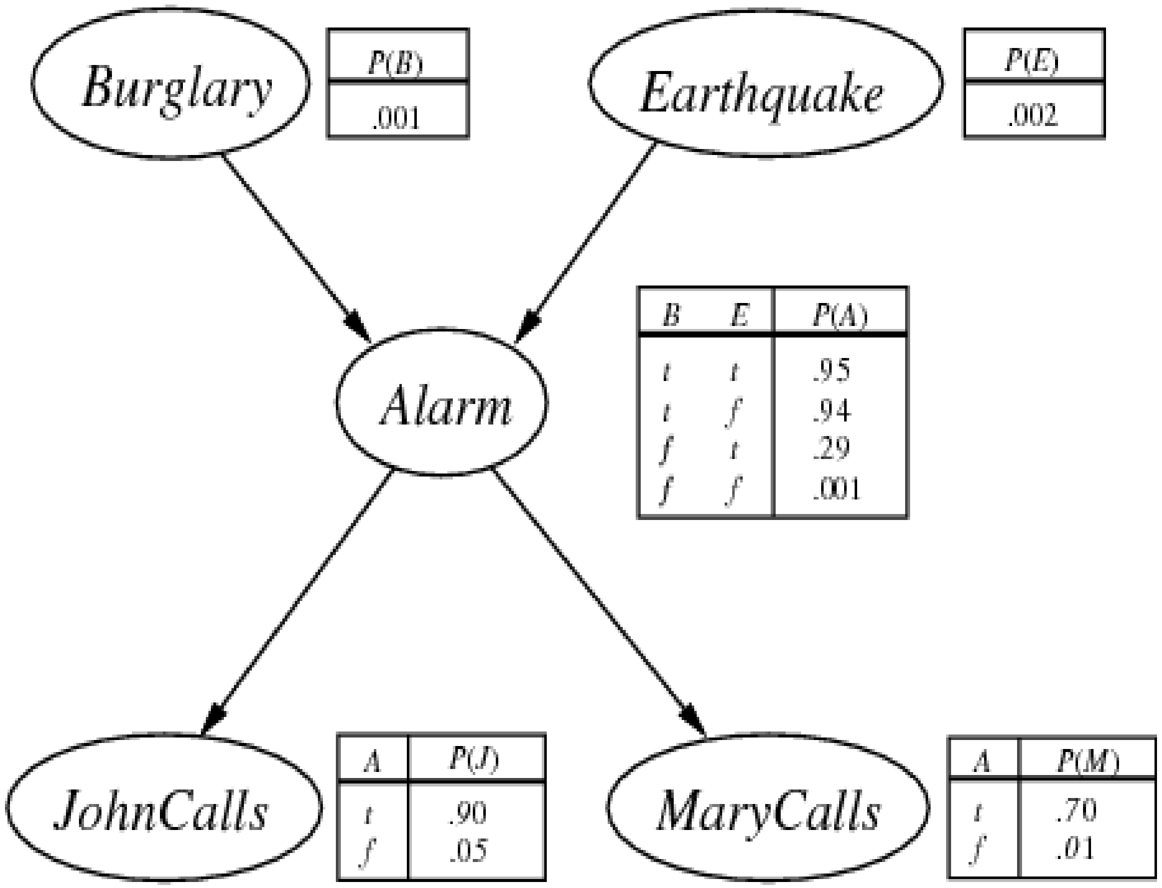


Bayesian Belief Network

- A Bayesian Belief Network is a method to describe the joint probability distribution of a set of variables.
- Let $x_1, x_2 \dots x_n$ be a set of random variables.
- A Bayesian Belief Network or BBN will tell us the probability of any combination of $x_1, x_2 \dots x_n$.

Bayesian Belief Network

- A BBN represents the joint probability distribution of a set of variables by explicitly indicating the assumptions of conditional independence through the following:
  1. Nodes representing random variables
  2. Directed links representing relations
  3. Conditional probability distributions
  4. The graph is a directed acyclic graph.
- Each variable is independent of its non-descendants given its predecessors. We say x_1 is a descendant of x_2 if there is a direct path from x_2 to x_1.



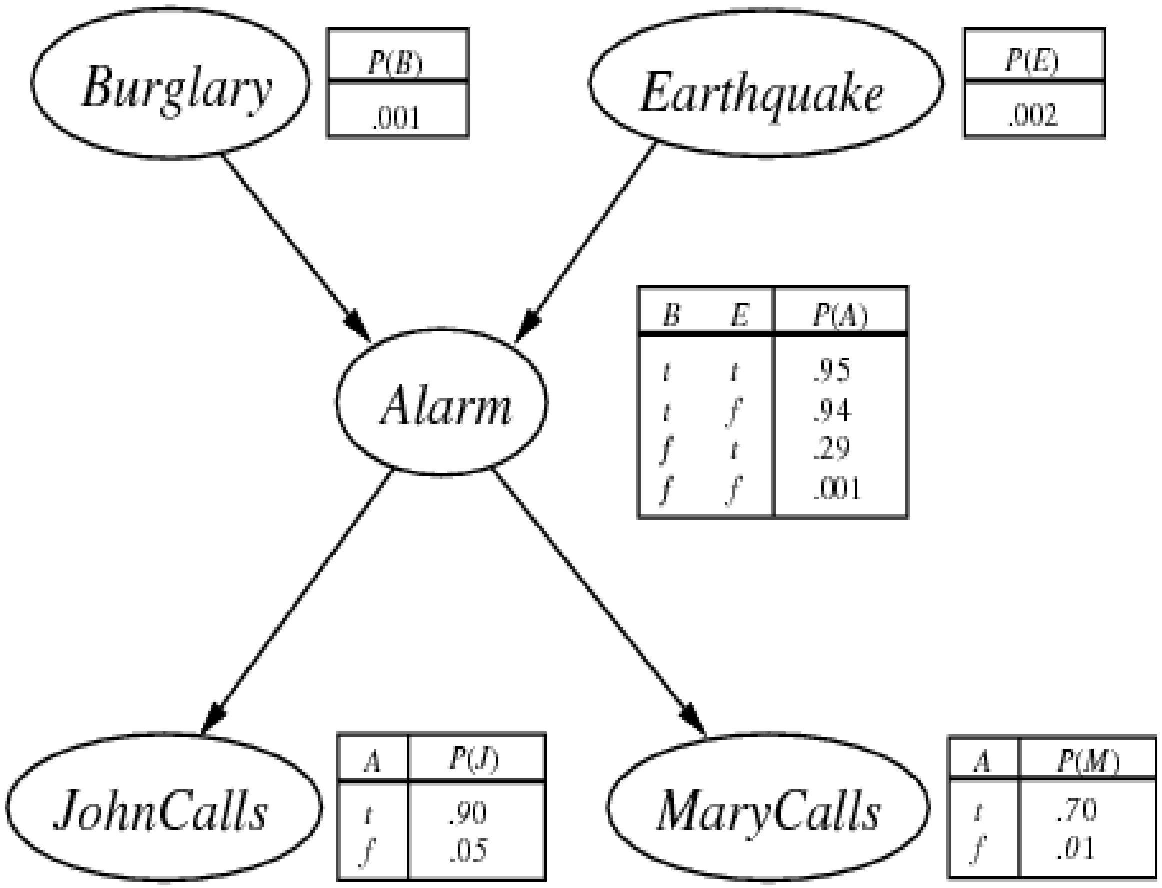
BBN: Joint Probability Distribution

- The joint probability distribution of a set of variables given a Bayesian Belief Network:

$$P(x_1, x_2 \dots x_n) = \prod P(x_i | Parents(x_i))$$

where parents are the immediate predecessors of x_i.

$$P(John, Mary, Alarm, \sim Bur, \sim EQ)$$
$$= P(John|Alarm)\, P(Mary|Alarm)$$
$$P(Alarm|\sim Bur, \sim EQ)\, P(\sim Bur) P(\sim EQ)$$



Questions?