

Basics of Probability for Machine Learning

Doug Downey, Northwestern EECS 349 Fall 2010

Probability in Machine Learning

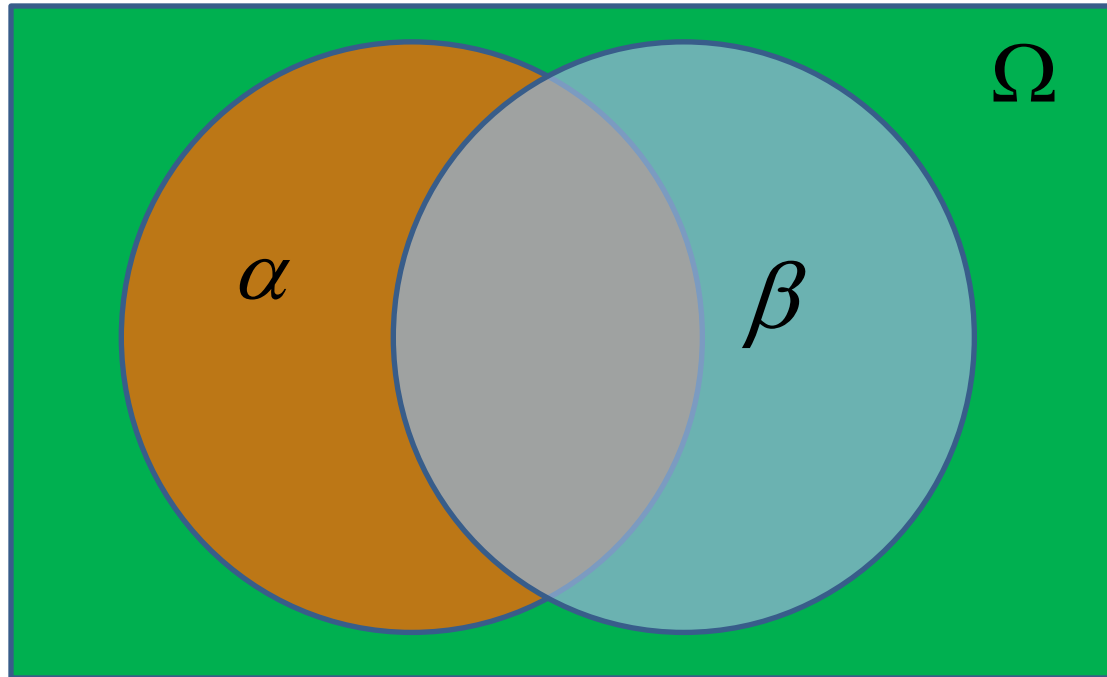
- The world is uncertain
 - Often want a probability distribution over possible outcomes
- Bayesian Methods in ML
 - Practical, widely used (e.g. Naïve Bayes)
 - Naturally incorporate prior knowledge

Events

- Event space Ω
 - E.g. for dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Set of measurable events $S \subseteq 2^\Omega$
 - E.g.,
 $\alpha = \text{event we roll an even number} = \{2, 4, 6\} \in S$
 - S must:
 - Contain the empty event \emptyset and the trivial event Ω
 - Be closed under union & complement
 - $\alpha, \beta \in S \rightarrow \alpha \cup \beta \in S$ and $\alpha \in S \rightarrow \Omega - \alpha \in S$



Probability Distributions



Can visualize probability as fraction of area

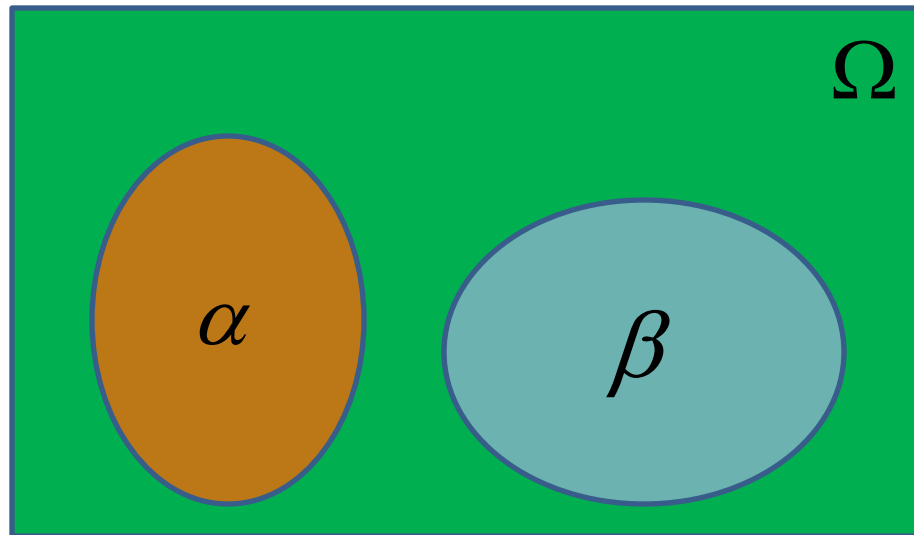
Probability Distributions

- A **probability distribution** P over (Ω, S) is a mapping from S to real values such that:

$$P(\alpha) \geq 0$$

$$P(\Omega) = 1$$

$$\alpha, \beta \in S \wedge \alpha \cap \beta = \emptyset \rightarrow P(\alpha \cup \beta) = P(\alpha) + P(\beta)$$



Probability: Interpretations & Motivation

- Interpretations
 - Frequentist
 - Bayesian/subjective
- Why use probability for subjective beliefs?
 - Beliefs that violate the axioms can lead to bad decisions *regardless* of the outcome [de Finetti, 1931]
 - Example: $P(A) = 0.6$, $P(\text{not } A) = 0.8$?
 - Example: $P(A) > P(B)$ and $P(B) > P(A)$?

Random Variables (1 of 2)

- A **random variable** is a function from Ω to a value
 - A short-hand for referring to *attributes* of events.
- E.g., your grade in this course
 - Let Ω = set of possible scores on hmwks and final
 - Cumbersome to have separate events GradeA, GradeB, GradeC
 - So instead define a random variable *Grade*
 - Deterministic function from Ω to {A, B, C}

Random Variables (2 of 2)

- Denote $P(\text{Grade} = A)$ as $P(\text{Grade} = A)$
 - Random variables will be in uppercase
 - When r.v. clear from context, abbreviate (e.g. $P(A)$)
- $\text{Val}(X)$ = set of values r.v. X can take
 - $\text{Val}(\text{Grade}) = \{A, B, C\}$
- Conjunction
 - Rather than write $P((\text{Grade} = A) \cap (\text{Age} = 21))$, we use $P(\text{Grade} = A, \text{Age} = 21)$ or just $P(A, 21)$.

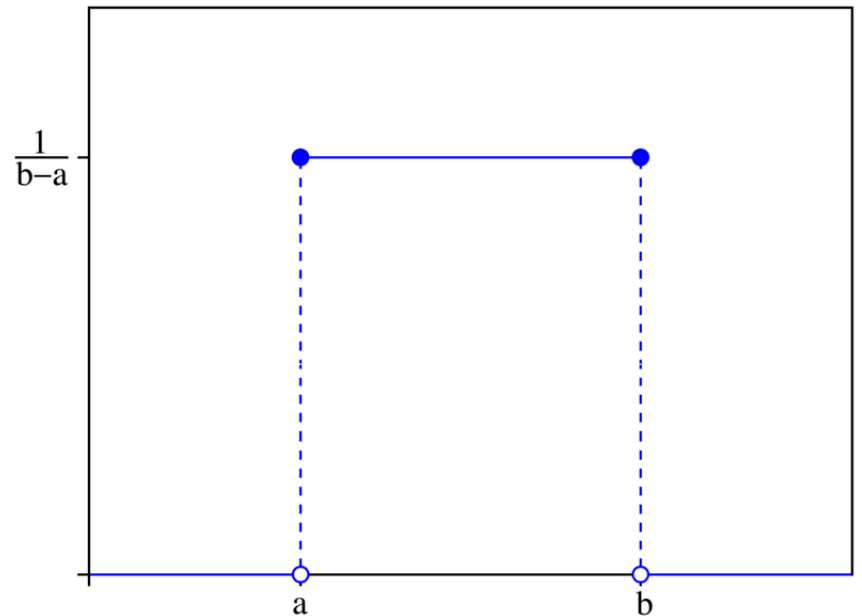
Continuous Random Variables

- For continuous r.v. X , specify a *density* $p(x)$, such that:

$$P(r \leq X \leq s) = \int_{x=r}^s p(x)dx$$

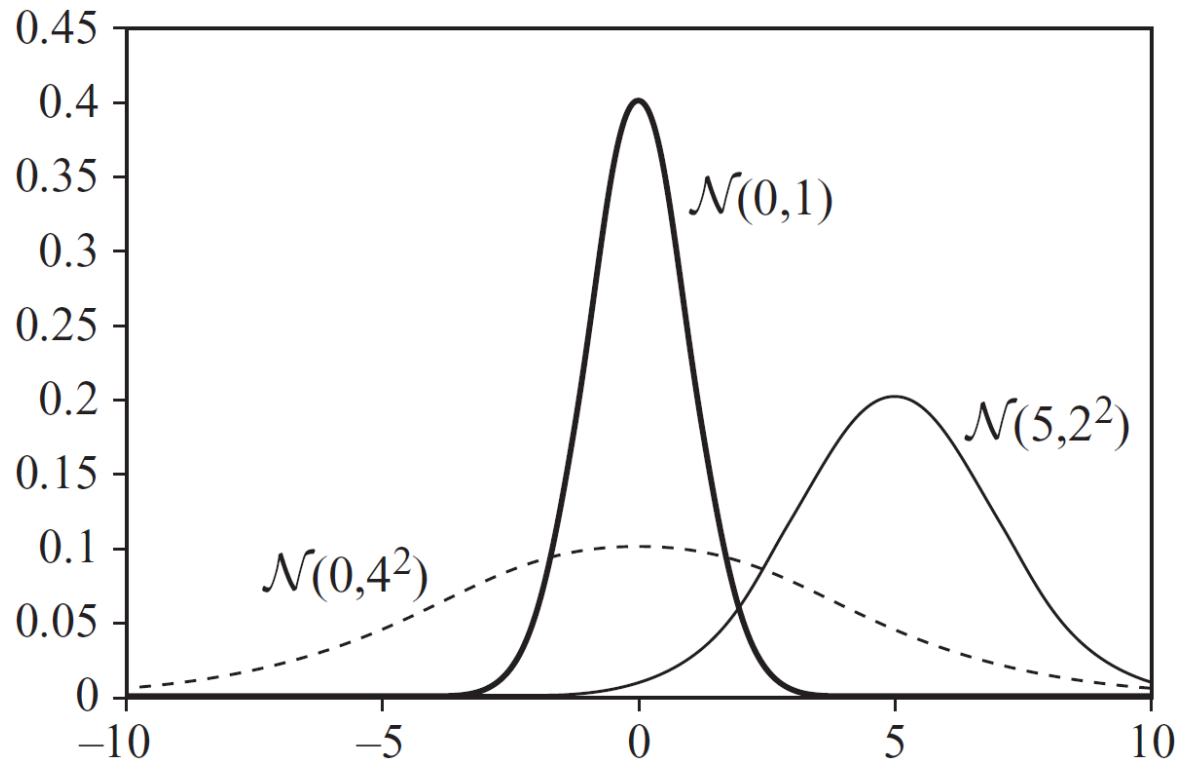
E.g.,

$$p(x) = \begin{cases} \frac{1}{a-b} & a \geq x \geq b \\ 0 & \text{otherwise} \end{cases}$$

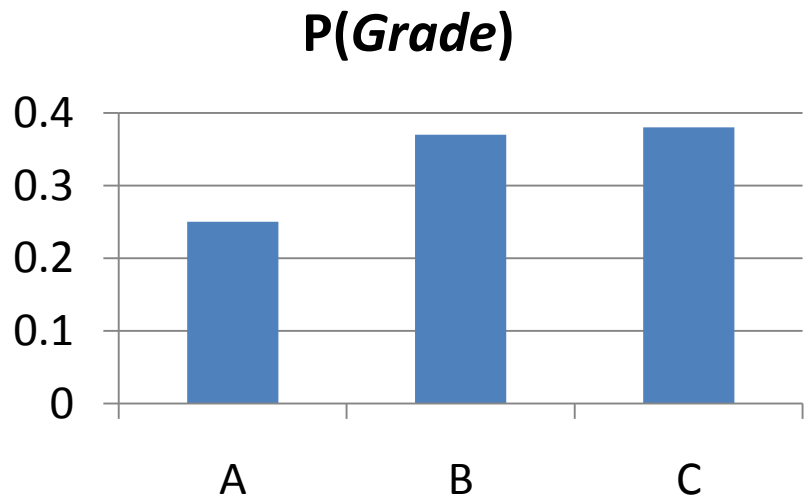
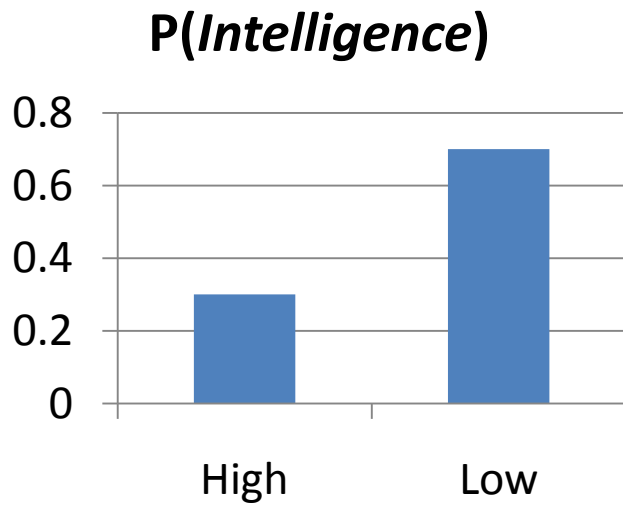


Gaussian Density

- $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$



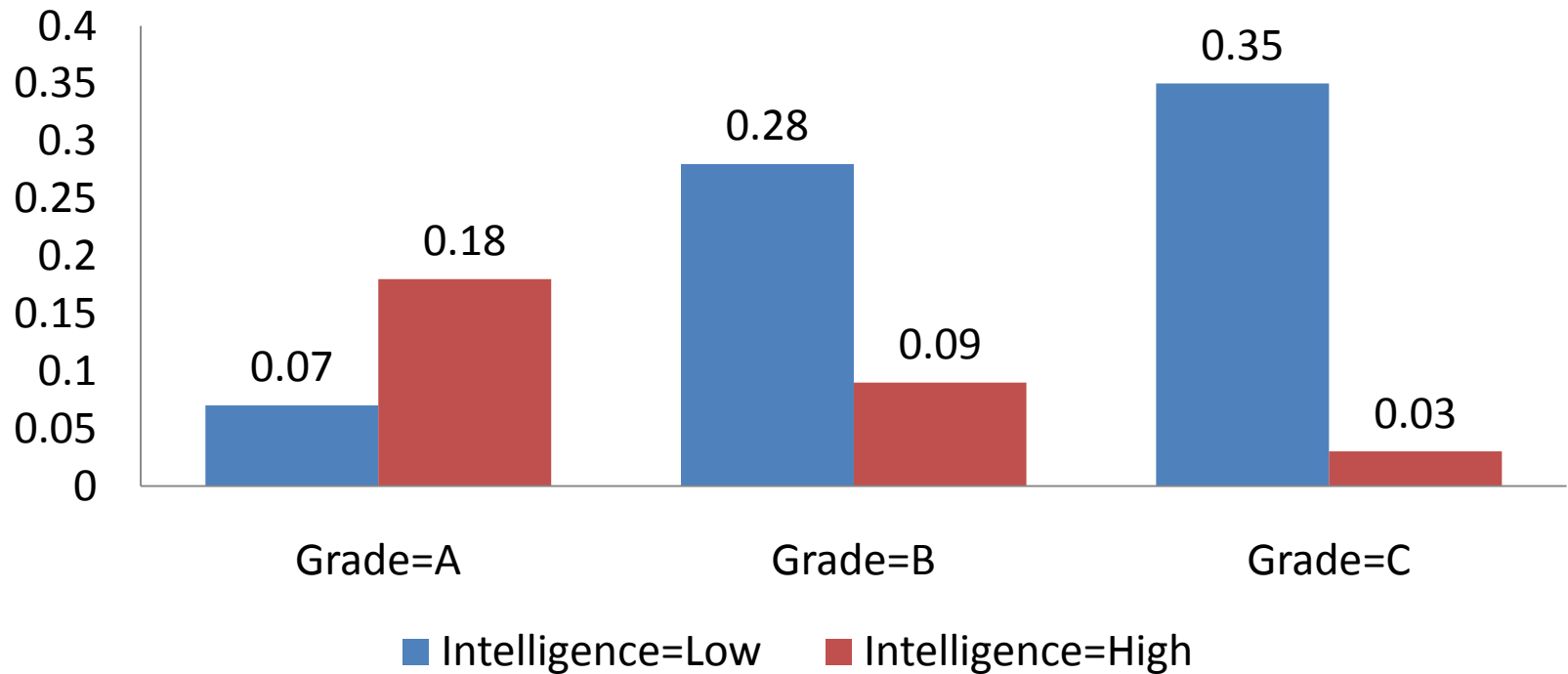
Distributions



- Called “marginal” because they apply to only one r.v.

Joint Distribution

P(Intelligence, Grade)



Joint Distribution

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

Joint Distribution specified with $2*3 - 1 = 5$ values

Joint Distribution

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

$P(\text{Grade} = \text{A}, \text{Intelligence} = \text{Low})?$ 0.07

Joint Distribution

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

$P(\text{Grade} = \text{A})? \quad 0.07 + 0.18 = 0.25$

Joint Distribution

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

$P(\text{Grade} = A \vee \text{Intelligence} = \text{High})?$

$$0.07 + 0.18 + 0.09 + 0.03 = 0.37$$

=> Given the joint distribution, we can compute probabilities for any proposition by summing events.

Conditional Probability

- $P(\text{Grade} = A \mid \text{Intelligence} = \text{High}) = 0.6$
 - the probability of getting an A given **only** *Intelligence* = High, and nothing else.
 - If we know *Motivation* = High or *OtherInterests* = Many, the probability of an A changes even given high *Intelligence*
- Formal Definition:
 - $P(\alpha \mid \beta) = P(\alpha, \beta) / P(\beta)$
 - When $P(\beta) > 0$

Conditional Probability

- Also:
 - $P(A \mid B, C) = P(A, B, C) / P(B, C)$
- More generally:
 - $P(\mathbf{A} \mid \mathbf{B}) = P(\mathbf{A}, \mathbf{B}) / P(\mathbf{B})$
 - (Boldface indicates vectors of variables)
- $P(\textit{Grade} = A \mid \textit{Grade} = A, \textit{Intelligence} = \textit{high})$?
- $P(\textit{CuriousGeorge} \mid \textit{MonkeyWithVacuum}, \textit{Cape})$?

Conditional Probability

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

$P(\text{Grade} = A \mid \text{Intelligence} = \text{High})$?

$P(\text{Grade} = A, \text{Intelligence} = \text{High}) = 0.18$

$P(\text{Intelligence} = \text{High}) = 0.18 + 0.09 + 0.03 = 0.30$

$\Rightarrow P(\text{Grade} = A \mid \text{Intelligence} = \text{High}) = 0.18 / 0.30 = \mathbf{0.6}$

Conditional Probability

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

$P(\text{Intelligence} \mid \text{Grade} = A)$?

Intelligence	
Low	High
0.28	0.72

Conditional Probability

		Intelligence	
		Low	High
Grade	A	0.28	0.72
	B	0.76	0.24
	C	0.92	0.08

$P(\text{Intelligence} \mid \text{Grade})?$

Actually three separate distributions, one for each *Grade* value

(has three independent parameters total)

Chain Rule

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i \mid X_{i-1} = x_{i-1}, \dots, X_1 = x_1)$$

- E.g., $P(\text{Grade}=\text{B}, \text{Int.} = \text{High})$
 $= P(\text{Grade}=\text{B} \mid \text{Int.} = \text{High})P(\text{Int.} = \text{High})$
- Can be used for distributions...
 - $P(A, B) = P(A \mid B)P(B)$

Handy Rules for Conditional Probability

- $P(A \mid B = b)$ is a single distribution, like $P(A)$
- $P(A \mid B)$ is *not* a single distribution
 - a *set* of $|\text{Val}(B)|$ distributions
- Any statement true for arbitrary distributions is also true if you condition on a new r.v.
 - $P(A, B) = P(A \mid B)P(B)$? (chain rule)
Then also $P(A, B \mid C) = P(A \mid B, C) P(B \mid C)$
- Likewise, any statement true for arbitrary distributions is also true if you replace an r.v. with two/more new r.v.s
 - $P(A \mid B) = P(A, B) / P(B)$? (def. of cond. Prob)
 - $P(A \mid C, D) = P(A, C, D) / P(C, D)$ or $P(\mathbf{A} \mid \mathbf{B}) = P(\mathbf{A}, \mathbf{B}) / P(\mathbf{B})$

Queries

- Given subsets of random variables Y and E , and assignments e to E
 - Find $P(Y \mid E = e)$
- Answering queries = **inference**
 - The whole point of probabilistic models, more or less
 - $P(\textit{Disease} \mid \textit{Symptoms})$
 - $P(\textit{StockMarketCrash} \mid \textit{RecentPriceActivity})$
 - $P(\textit{CodingRegion} \mid \textit{DNASequence})$
 - $P(\textit{PlayTennis} \mid \textit{Weather})$
 - ...**(the other key task is learning)**

Answering Queries: Summing Out

		Intelligence = Low		Intelligence=High	
		Time=Lots	Time=Little	Time=Lots	Time=Little
Grade	A	0.05	0.02	0.15	0.03
	B	0.14	0.14	0.05	0.0
	C	0.10	0.25	0.01	0.02

$P(\text{Grade} \mid \text{Time} = \text{Lots})?$

$$\sum_{v \in \text{Val}(\text{Intelligence})} P(\text{Grade}, \text{Intelligence} = v \mid \text{Time} = \text{Lots})$$

MAP Queries

- Given subsets of random variables \mathbf{Y} and \mathbf{E} , and assignments \mathbf{e} to \mathbf{E}
 - Find $\text{MAP}(\mathbf{Y} \mid \mathbf{e}) = \arg \max_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{e})$
- MAP stands for “maximum a posteriori”
 - (more later)

Answering Queries: Solved?

- Given the joint distribution, we can answer any query by summing
- ...but, joint distribution of 500 Boolean variables has $2^{500} - 1$ parameters (about 10^{150})
- For non-trivial problems (~ 25 boolean r.v.s or more), using the joint distribution requires
 - Way too much **computation** to compute the sum
 - Way too many **observations** to learn the parameters
 - Way too much **space** to store the joint distribution

Conditional Independence (1 of 3)

- Independence
 - $P(A, B) = P(A) * P(B)$, denoted $A \perp B$
 - E.g. consecutive dice rolls
 - Gambler's fallacy
 - Rare in (real) applications



Conditional Independence (2 of 3)

- Conditional Independence
 - $P(A, B | C) = P(A | C) P(B | C)$, denoted $(A \perp B | C)$
 - Much more common
 - E.g.,
(GetIntoNU \perp GetIntoStanford | Application),
but **NOT** *(GetIntoNU \perp GetIntoStanford)*



Conditional Independence (3 of 3)

- How does Conditional Independence save the day?

$$P(NU, Stanford, App) =$$

$$P(NU | Stanford, App) * P(Stanford | App) * P(App)$$

Now, $(A \perp B | C)$ means $P(A | B, C) = P(A | C)$

So since $(NU \perp Stanford | App)$, we have

$$P(NU, Stanford, App) =$$

$$P(NU | App) * P(Stanford | App) * P(App)$$

Say $Val(App) = \{Good, Bad\}$ and $Val(School) = \{Yes, No, Wait\}$

All we need is $4+4+1=9$ numbers

(vs. $3*3*2-1=17$ for the full joint)

- Full joint has size **exponential** in # of r.v.s
Conditional independence eliminates this!



Properties of Conditional Independence

- Decomposition
 - $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z)$
- Weak Union
 - $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z, W)$
- Contraction
 - $(X \perp W \mid Z, Y) \& (X \perp Y \mid Z) \Rightarrow (X \perp Y, W \mid Z)$

Bayes' Rule

- $P(A | B) = P(B | A) P(A) / P(B)$

- Example:

$$P(\text{symptom} | \text{disease}) = 0.95, P(\text{symptom} | \neg\text{disease}) = 0.05$$

$$P(\text{disease}) = 0.0001$$

$$P(\text{disease} | \text{symptom})$$

$$= \frac{P(\text{symptom} | \text{disease}) * P(\text{disease})}{P(\text{symptom})}$$

$$= \frac{0.95 * 0.0001}{0.95 * 0.0001 + 0.05 * 0.9999} = \mathbf{0.002}$$

Terms for Bayes

$$P(\textit{Model} | \textit{Data}) = \frac{P(\textit{Data} | \textit{Model}) P(\textit{Model})}{P(\textit{Data})}$$

$P(\textit{Model})$: **Prior**

$P(\textit{Data} | \textit{Model})$: **Likelihood**

$P(\textit{Model} | \textit{Data})$: **Posterior**

What have we learned?

- Probability – a calculus for dealing with uncertainty
 - Built from small set of axioms (ignore at your peril)
- Joint Distribution $P(A, B, C, \dots)$
 - Specifies probability of all combinations of r.v.s
 - Intractable to compute exhaustively for non-trivial problems
- Conditional Probability $P(A \mid B)$
 - Specifies probability of A given B
- Conditional Independence
 - Can radically reduce number of variable combinations we must assign unique probabilities to.
- Bayes' Rule