Basics of Probability for Machine Learning

Doug Downey
Northwestern EECS 349 Winter 2014

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

Random Variables (1 of 2)

- A random variable is a variable whose value is the result of an experiment
  - A short-hand for referring to attributes of **events**
- E.g., your grade in this course
  - Atomic Events Ω = set of possible scores on hmwks and final
  - Define a random variable *Grade*
    - Deterministic function from Ω to {A, B, C}

Random Variables (2 of 2)

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

Probability: Interpretations & Motivation

- Probabilities assign numbers to events
- Axioms
  - P(a) >= 0
  - P(union of all possible events) = 1
  - When events $\alpha \cap \beta = \varnothing$ then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$
- 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]

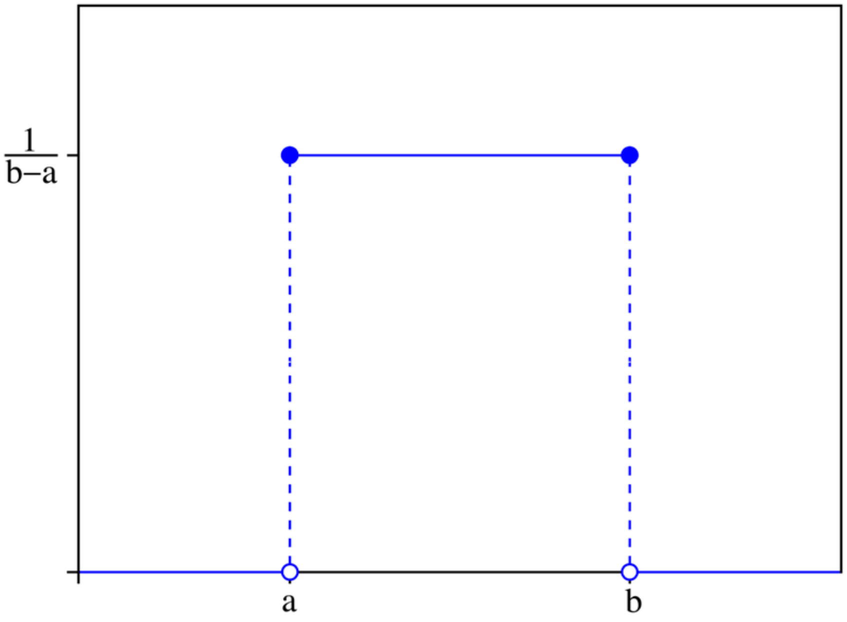
Continuous Random Variables

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

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

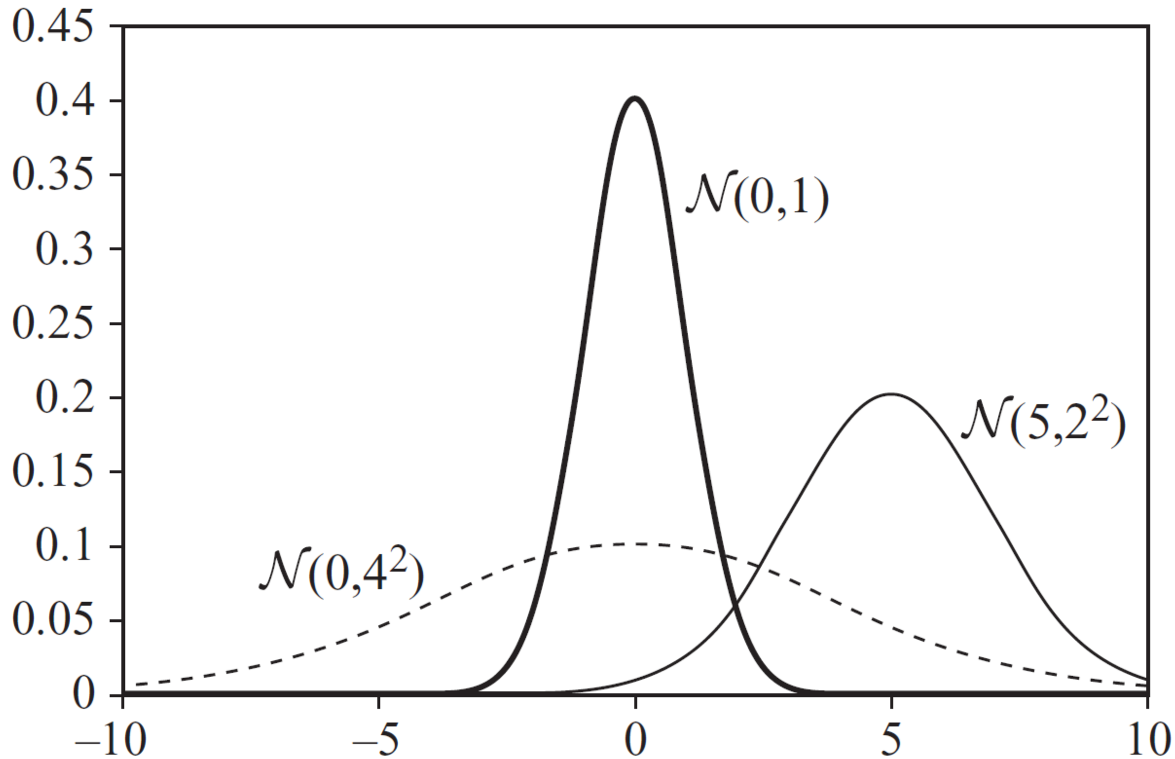
E.g.,

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

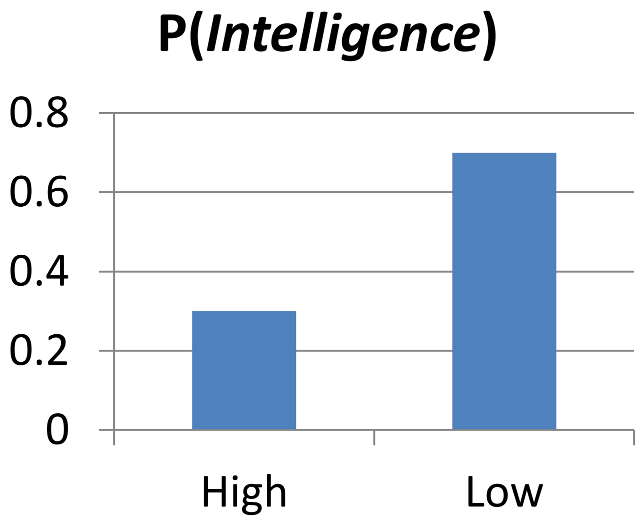


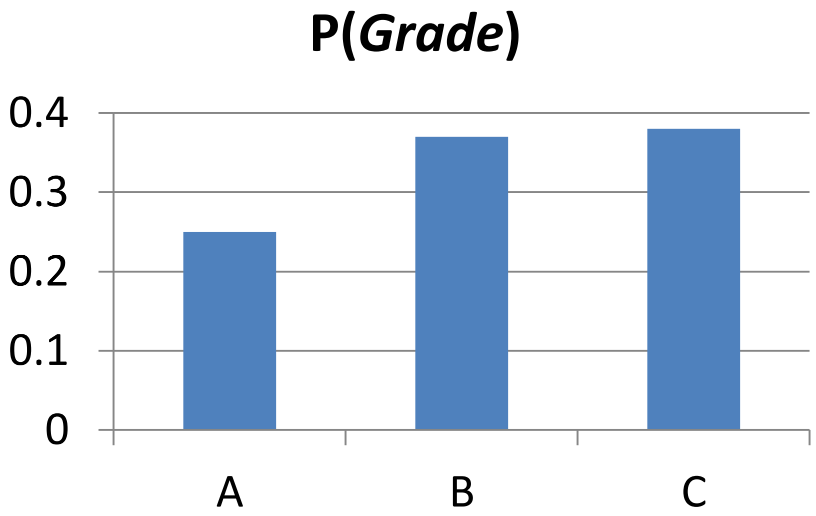
Gaussian Density

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



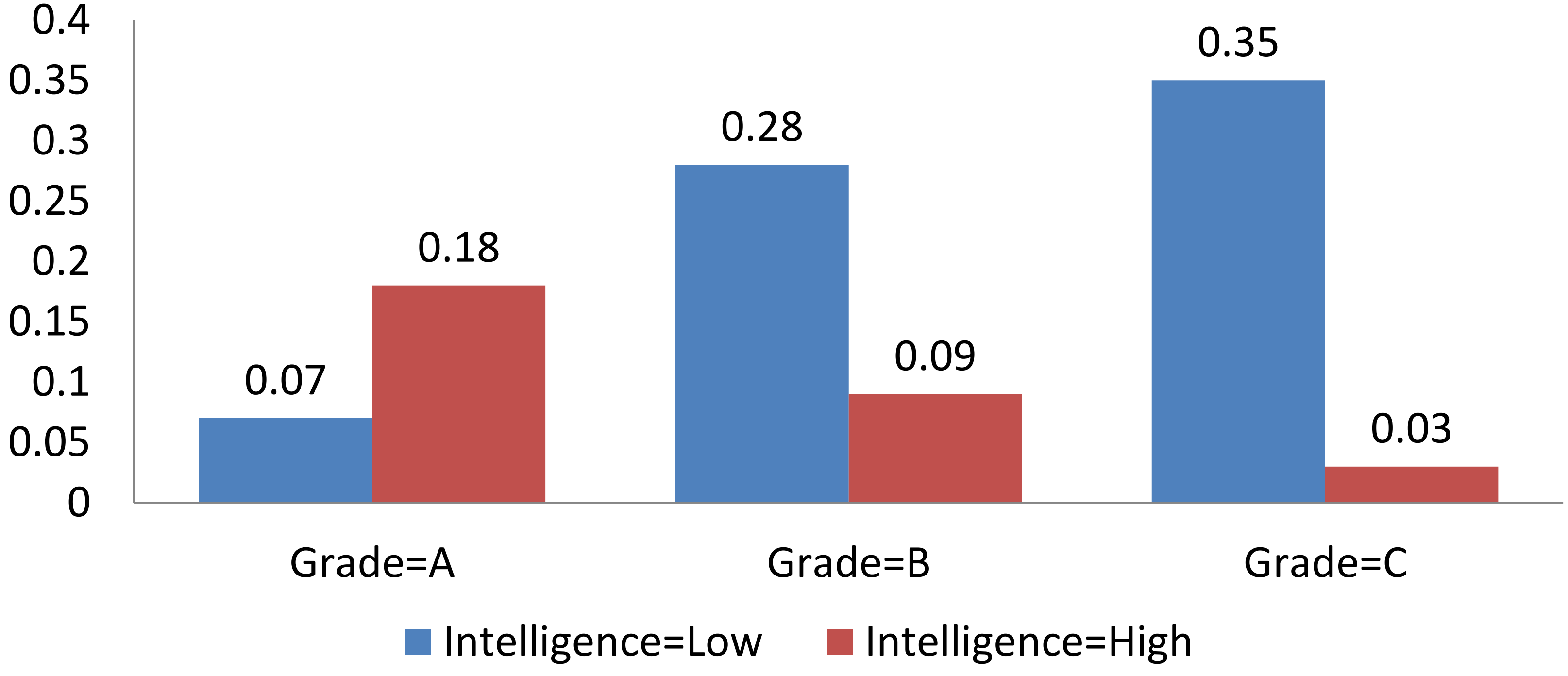
Distributions





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(Grade = A, Intelligence = 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(Grade = A)? 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(*Grade* = A | *Intelligence* = 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:
  - $P(\alpha \mid \beta) = \mathrm{P}(\alpha, \beta) / \mathrm{P}(\beta)$
    - When $\mathrm{P}(\beta) > 0$

Conditional Probability

- Also:
  - P(A | B, C) = P(A, B, C) / P(B, C)

- More generally:
  - P(***A*** | ***B***) = P(***A, B***) / P(***B***)
  - (Boldface indicates vectors of variables)

- P(*Grade* = A | *Grade* = A, *Intelligence* = high) ?
- P(*CuriousGeorge* | *MonkeyWithVacuum*, *Cape*)?

Conditional Probability

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

P(*Grade* = A | *Intelligence* = High) ?

P(*Grade* = A, *Intelligence* = High) = 0.18

P(*Intelligence* = High) = 0.18+0.09+0.03 = 0.30

=> P(*Grade* = A | *Intelligence* = High) = 0.18/0.30 = **0.6**

Conditional Probability

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

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

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(*Intelligence* | *Grade*)?

Actually three separate distributions, one for each *Grade* value
(has three independent parameters total)

Chain Rule

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

- E.g., P(*Grade*=B, *Int.* = High)
  = P(*Grade*=B | *Int.*= High)P(*Int.* = High)
- Can be used for distributions…
  - P(*A*, *B*) = P(*A* | *B*)P(*B*)

Handy Rules for Conditional Probability

- P(A | B = b) is a single distribution, like P(A)
- P(A | B) is *not* a single distribution
  - a *set* of |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 | B)P(B)? (chain rule)
    Then also P(A, B | C) = P(A | B, C) P(B | 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 | B) = P(A, B) / P(B) ? (def. of cond. Prob)
  - P(A | C, D) = P(A, C, D) / P(C, D) or P($\boldsymbol{A}$ | $\boldsymbol{B}$) = P($\boldsymbol{A}$, $\boldsymbol{B}$) / P($\boldsymbol{B}$)

Queries

- Given subsets of random variables ***Y*** and ***E***, and assignments ***e*** to ***E***
  - Find P(***Y*** | ***E* = *e***)
- Answering queries = **inference**
  - The whole point of probabilistic models, more or less
  - P(*Disease* | *Symptoms*)
  - P(*StockMarketCrash* | *RecentPriceActivity*)
  - P(*CodingRegion* | *DNASequence*)
  - …(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

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

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

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, ($\boldsymbol{A} \perp \boldsymbol{B} \mid \boldsymbol{C}$) means P($\boldsymbol{A} \mid \boldsymbol{B}, \boldsymbol{C}$) = P($\boldsymbol{A} \mid \boldsymbol{C}$)

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

  P(*NU, Stanford, App*) =

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

  Say $App \in$ {Good, Bad} and $School \in$ {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!

Bayes' Rule

- $P(A \mid B) = P(B \mid A)\, P(A) \,/\, P(B)$
- Example:

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

  P(disease = 0.0001)

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

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

Bayes' Rule

- $P(A \mid B) = P(B \mid A)\, P(A) \,/\, P(B)$
- Also:
  - $P(A \mid B, C) = P(B \mid A, C)\, P(A \mid C) \,/\, P(B \mid C)$
- More generally:
  - $P(\boldsymbol{A} \mid \boldsymbol{B}) = P(\boldsymbol{B} \mid \boldsymbol{A})\, P(\boldsymbol{A}) \,/\, P(\boldsymbol{B})$
  - (Boldface indicates vectors of variables)

Terms for Bayes

P(*Model*| *Data*) =
P(*Data* | *Model*) P(*Model*) / P(*Data*)

P(*Model*) : **Prior**

P(*Data* | *Model*) : **Likelihood**

P(*Model* | *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, …)
  - Specifies probability of all combinations of r.v.s
  - Intractable to compute exhaustively for non-trivial problems
- Conditional Probability P(A | B)
  - Specifies probability of A given B
- Conditional Independence
  - Can radically reduce number of variable combinations we must assign unique probabilities to.
- Bayes' Rule