Basic Probability Review

CPS 271
Ron Parr

Probability: Who needs it?

- Learning without probabilities is possible
  - Version spaces
  - Explanation based learning

- Learning almost always involves
  - Noise in data
  - Prediction about the future

- Learning systems that don’t use probability in some way tend to be very, very brittle
Probabilities

- Natural way to represent uncertainty
- People have intuitive notions about probabilities
- Many of these are wrong or inconsistent
- Most people don’t get what probabilities mean
- Finer details of this question still debated

Relative Frequencies

- Probabilities defined over events
- Space of all possible events is “event space”

Think: Playing blindfolded darts with the Venn diagram..
Understanding Probabilities

• Probabilities have dual meanings
  – Relative frequencies (frequentist view)
  – Degree of belief (Bayesian/subjectivist view)

• Neither is entirely satisfying
  – No two events are truly the same
    (reference class problem)
  – Statements should be grounded in reality in some way

Why probabilities are good
(despite the difficulties)

• Subjectivists: probabilities are degrees of belief

• Are all degrees of belief probability?
  – AI has used many notions of belief:
    • Certainty Factors
    • Fuzzy Logic

• Can prove that a person who holds a system of beliefs inconsistent with probability theory can be tricked into accepted a sequence of bets that is guaranteed to lose
Probabilities over discrete events
(and the horror of common notation)

- Probabilities defined over sets of random variables
- RVs usually represented with capitals: $X,Y,Z$
- Use lower case letters for values from domains
- $X=x$ asserts that the random variable $X$ has taken on value $x$
- $P(x)$ is shorthand for $P(X=x)$

Event spaces for discrete RVs

- 2 variable case

```
  \[ \begin{array}{c|c}
  ab & ab \\
  \hline
  \bar{a}b & \bar{a}b \\
  \end{array} \]
```

- Important: Event space grows exponentially in number of random variables
- Components of event space = atomic events
Joint Distributions

- A joint distribution is an assignment of probabilities to every possible atomic event
- We can define all other probabilities in terms of the joint probabilities:

\[
P(a) = \sum_{e_i \in e(a)} P(e_i)
\]

\[
P(a) = P(ab) + P(a\overline{b})
\]

- AKA: Sum rule, marginalization

Why Probabilities Are Messy

- Probabilities are not truth-functional

- To compute \(P(a \text{ and } b)\), need joint distribution
  - sum out all of the other events from distribution
  - In general, it is not a function of \(P(a)\) and \(P(b)\)
  - In general, it is not a function of \(P(a)\) and \(P(b)\)
  - In general, it is not a function of \(P(a)\) and \(P(b)\)
  - This fact led to many approximations methods such as certainty factors and fuzzy logic (Why?)
Independence

- RVs A and B are independent iff:
  - \( P(AB) = P(A)P(B) \)
- Independence:
  - Make things computationally easy
  - Makes things boring
    - From an algorithmic standpoint
    - From a predictive standpoint
  - Is almost never true
  - Is approximately true for “unrelated” events

Kolmogorov’s axioms of probability

- \( 0 \leq P(a) \leq 1 \)
- \( P(\text{true}) = 1; P(\text{false}) = 0 \)
- \( P(a \text{ or } b) = P(a) + P(b) - P(ab) \)
  - Subtract to correct for double counting

- This is sufficient to specify probability theory for discrete variables
- Continuous variables need density functions
Continuous Random Variables

- Domain is some interval, region, or union of regions
- Uniform case: Simplest to visualize (event probability is proportional to area)
- Non-uniform case visualized with extra dimension

Gaussian (normal/bell distribution:)

Requirements on Continuous Distributions

- $p(x)>1$ is possible so long as:
  \[ \int_x p(x)dx = 1 \]
- Don’t confuse $p(x)$ and $P(X=x)$
- $P(X=x)$ for any $x$ is 0!
  \[ P(x \in A) = \int_A p(x)dx \]
Cumulative Distributions

- When distribution is over numbers, we can ask:
  - $P(X \geq c)$ for some $c$
  - $P(X < c)$ for some $c$
  - $P(a \leq X \leq b)$ for some, $a$ and $b$
- Solve by
  - Summation
  - Integration
- Cumulative sometimes called
  - CDF
  - Distribution function

Sloppy Comment about Continuous Distributions

- In many, many cases, you can generalize what you know about discrete distributions to continuous distributions, replacing “$P$” with “$p$” and “$\Sigma$” with “$\int$”
- Proper treatment of this topic requires measure theory and is beyond the scope of the text and class
Conditional Probabilities

- Ordinary probabilities = unconditional or prior probabilities

- \( P(a \mid b) = \) probability of a given that we know only \( b \)

- If we know \( c \) and \( d \), we can’t use \( P(a \mid b) \) directly (annoying, but important detail!)

- \( P(a \mid a) = 1 \)

Conditional Probability

- \( P(b \mid a) = \) Probability of event \( b \) given that event \( a \) is true

- Idea: In what fraction of a event space is \( b \) also true?

\[
P(B \mid A) = \frac{P(AB)}{P(A)}
\]
Definition of Conditional Probability

- Following geometric intuitions from previous slide
  - \( P(B | A) = \frac{P(AB)}{P(A)} \)
  - \( P(A | B) = \frac{P(AB)}{P(B)} \)
- Also known as the **product rule**:
  - \( P(B | A)P(A) = P(AB) = P(BA) \)
  - \( P(A | B)P(B) = P(AB) = P(BA) \)

Condition with Bayes’s Rule

\[
P(AB) = P(BA) \\
P(A | B)P(B) = P(B | A)P(A) \\
P(A | B) = \frac{P(B | A)P(A)}{P(B)}
\]
Why Bayes’s Rule is Cool

- Solves the “inverse probability” problem
- Diagnosis:
  - Often we know: $P(\text{Symptoms} | \text{Disease})$ from data
  - Want: $P(\text{D} | S)$ to diagnose patients
- Sensing:
  - Know: $P(\text{Observation} | \text{Reality})$
  - Want: $P(R | O)$
- Learning:
  - Know: $P(\text{Data} | \text{Hypothesis about source model})$
  - Want: $P(H | D)$

Expectation

\[ E(X) = \sum_x xP(x) \]

- Matches some colloquial notions of average
- “Mean”
- Arithmetic mean (uniform weights)
- For continuous random variables:

\[ E(X) = \int_x xp(x)dx \]

Nota bene: We will be assuming that $E(X)$ is finite.
Properties of Expectation

\[ E(f(X)) = \sum_x f(x)p(x) \]

\[ E(aX) = ???, \quad aE(X) \]
\[ E(aX + b) = ???, \quad aE(X) + b \]
\[ E(X + Y) = ???, \quad E(X) + E(Y) \]
\[ E(XY) = ???, \quad \text{If } X, Y \text{ are independent: } E(X)E(Y) \]

Variance

- Expected, squared deviation from the mean
- “How much we trust the mean”

\[ Var(X) = E[(X - E(X))^2] \]
\[ = E(X^2) - E(X)^2 \]

Nota bene: We will typically assume that \( Var(X) \) is finite.
Properties of Variance

\[ \sigma^2(X) = Var(X) = E[(X - E(X))^2] \]

\[ Var(aX) = ??? \quad a^2 Var(X) \]
\[ Var(aX + b) = ??? \quad a^2 Var(X) \]
\[ Var(X + Y) = ??? \quad Var(X) + Var(Y) + 2E[(X - E(X))(Y - E(Y))] \]

If X,Y are independent: \[ Var(X) + Var(Y) \]

Covariance

\[ Var(X + Y) = Var(X) + Var(Y) + 2E[(X - E(X))(Y - E(Y))] \]

Covariance captures the leftover:

\[ Cov(X,Y) = Cov(Y,X) = E[(X - E(X))(Y - E(Y))] \]
\[ Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y) \]

If X,Y are independent, \[ cov(X,Y) = 0 \]
Standard Deviation

\[ \sigma(X) = SD(X) = \sqrt{Var(X)} \]

- Expected deviation from the mean
- Sometimes more natural than variance:
  \[ \sigma(aX) = a \sigma(X) \]
- Often not, for X,Y independent:
  \[ \sigma(X + Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)} \]

Sample (Empirical) Mean

- Suppose we observe \( X_1 \ldots X_n \)
- Assume these are independently drawn, and indentically distributed (IID)

  \[ \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \]

- What is our estimate for \( E(X) \)?

- Why?
  \[ E(\bar{X}) = E\left( \frac{\sum_{i=1}^{n} X_i}{n} \right) = \frac{nE(X)}{n} = E(X) \]

Also...
Properties of the Sample Mean

- Is unbiased (shown on previous slide) and consistent (shown later)
- Minimizes expected squared loss on the training data
- Is the maximum likelihood (ML) estimate of the true mean given a variety of benign modeling assumptions

Sample Variance

- Suppose we observe \(X_1 \ldots X_n\)
- Assume these are independently drawn, and identically distributed (IID)
- What is our estimate for \(V(X)\)?

\[
\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} (\bar{X} - X_i)^2}{n-1}
\]

- Why?
Chebyshev’s Inequality

• Let $X$ have finite mean and variance:

$$P(|X - E(X)| \geq c) \leq \frac{\text{Var}(X)}{c^2}$$

• For constant $c$:
  – Decreasing variance decreases prob. of being far from mean

• For constant variance
  – Prob. of missing a +/- $c$ interval drops quickly with $c$

• Note: No distribution assumptions

Convergence of Sample Mean

• Apply Chebyshev’s inequality to sample mean

$$P\left(\left|\bar{X} - E(X)\right| \geq c\right) \leq \frac{\text{Var}(\bar{X})}{c^2}$$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \sum_{i=1}^{n} \frac{1}{n^2} \text{Var}(X_i) = \frac{\text{Var}(X)}{n}$$

$$\lim_{n \to \infty} P\left(\left|\bar{X} - E(X)\right| \geq c\right) \leq \frac{\text{Var}(X)}{nc^2} = 0$$