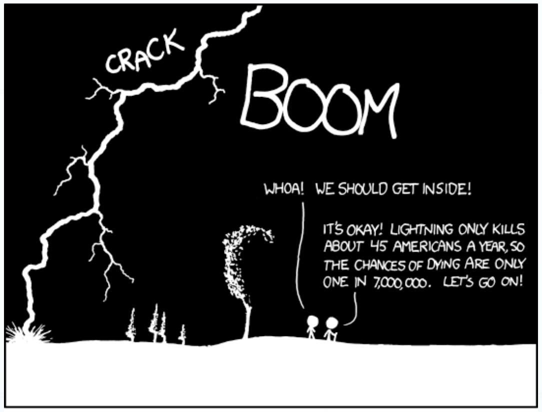


THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Unsupervised Learning: Probability

Thanks to Hal Daume for some content

Review: Supervised Learning

- Supervised methods
  - Focus on constructing objective functions
    - Different classifiers (approaches) all trying to solve the same problem
  - Some exceptions (e.g. boosting)

Unsupervised Learning

- Focus on a single objective function (likelihood)
- Cover different frameworks and methods
  - Clustering: outline goals of unsupervised learning and basic approaches
  - EM algorithm: how we learn without labeled data
  - Graphical models: how we construct models
  - Structured prediction: construct models with different output types
  - Manifold learning/dimensionality reduction

Today: Probability

- Probability has shown up in supervised learning
  - Maximize data likelihood
    - Logistic regression
    - Least squares regression
- The language of how we express models
- Today: review core concepts

Probability is Hard

- Probability is not intuitive
- People routinely make wrong decisions
- Examples
  - The lottery
  - Streaks in sports
  - 50/50 chances
  - http://www.cc.com/video-clips/hzqmb9/the-daily-show-with-jon-stewart-large-hadron-collider
    - 2:05-3:36

Probability 101

- Probability theory assigns a numerical probability to events
- Probability of an event = *fraction of times* that event would occur if we ran an experiment many times
  - This is the *frequentist* definition of probability
- Events are drawn (they happen) from a sample space Ω (omega)
- A probability model is a function that maps any subset of Ω to a real value between 0 and 1
- Formally, $p : P(\Omega) \rightarrow [0, 1]$

3 Axioms of Probability

- $p(E) \geq 0$ for all $E \subseteq \Omega$
  - Cannot have a negative event
- $p(\Omega) = 1$
  - An event must occur
- If $E_1, E_2, E_3, \cdots \subseteq \Omega$, are pairwise disjoint, then: $p(E_1 \cup E_2 \cup E_3 \cup \ldots) = p(E_1) + p(E_2) + p(E_3) + \ldots$
  - Events are additive

Rules of Probability

- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
  - Decomposition of "or"

Joint Conditional

- $p(\Omega \setminus E) = 1 - p(E)$
  - Inclusion/exclusion
- $p(A \cap B) = p(A,B) = p(A|B)\, p(B)$
  - Decomposition of "and", called the product rule
- $p(\varnothing) = 0$
  - Null event

Why We Need These Rules

- Many machine learning models are probabilistic
  - Assume data fits some probability distribution
  - Example: logistic regression
    - The label is given by a conditional probability based on the data $p(y|x)$
  - Alternatively: define a joint probability of label and data $p(y,x)$

$$p(y,x) = p(y|x)p(x) = p(x|y)p(y)$$

    - We can model these separately
  - What is $p(y=0|x)$

$$p(y=0|x) = 1 - p(y=1|x)$$

Expectations

- We assign values to these probabilities given data
- Using our learned probabilities, we want to make predictions about what we **expect** will happen
  - The "average" result of an event
- If A is a random variable, the expectation is

$$\mathrm{E}_p[A] = \begin{cases} \sum_{a \in A} p(A=a)a & \text{discrete} \\ \int_A da\, p(A=a)a & \text{continuous} \end{cases}$$

Bayes Rule

- Sometimes we cannot measure p(y|x) directly
  - Ex. y is a disease and x is a symptom
- But we can measure p(x|y)

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Medical Test

- You take a test for a disease. The results are positive.
- The probability that the test shows positive given a negative sample is 10,000 to 1
- Do you have the disease?

Applications of Bayes Rule

- *p*(test=positive|disease=negative) = 0.0001
- *p*(test=negative|disease=negative) =0.9999
- *p*(test=positive|disease=positive)=0.9999
- *p*(disease=negative)=0.999999

- *p*(disease=negative|test=positive) = p(test=positive|disease=negative) p(disease=negative)/p(test=positive) = 9x10-5 * K
- *p*(disease=positive|test=positive) = p(test=positive|disease=positive) p(disease=positive)/p(test=positive) = 9x10-7*K
- (9x10-5 + 9x10-7)* K = 1
- *p*(disease=positive|test=positive) ≈ 0.01 (1%)

Chain Rule

- We want to know the probability of many things happening
- $p(x_1, x_2, x_3, \ldots x_n)$
  - Often difficult to manage a large joint probability
- Break apart using the chain rule

$$p(x_1, x_2, x_3, \ldots x_N) = \prod_{n=1}^{N} p(x_n | x_1, x_2 \ldots x_{n-1})$$

Why the Chain Rule?

- Assume you have many events you want to predict
  - e.x., probability of an image- prediction for each pixel
- P(image) = P(pixel_1, pixel_2, pixel_3, ...)
  - This is a very large probability
  - Hard to make good estimates
    - Most of the time it will be 0
- Break down using the chain rule
  - P(image) = P(pixel_n | pixel_1, pixel_2...) *
    P(pixel_n-1 | pixel_1, pixel_2...)

  This would be useful if we could remove terms on the RHS of the probability...

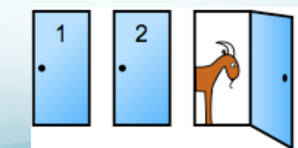
Marginalization

- We want to compute *p*(x)
- But we only know p(x|y)
- Solution: sum over all possible y

$$p(X = x) = \sum_{y} p(X = x | Y = y) p(Y = y)$$

Monty Hall: A Game Show

- Three doors: behind 1 door is a car, the other two have goats
  - You want the car
- You pick a door
- The host opens a different door and reveals a goat
  - He knows where the car is
- Should you switch doors?



Simulation

Probability Distributions

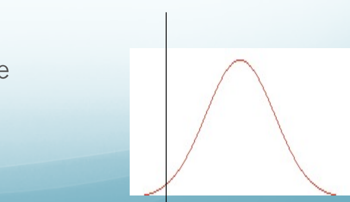
- Function that maps input to an output such that
  - Output must be between 0 and 1
  - Total area under the function must be 1

Gaussian Distribution

- Gaussian distribution (aka. normal distribution)

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

- μ- mean
- σ- standard deviation, σ^2 variance
- Continuous distribution (assigns probabilities to real valued numbers)
- Many special properties that will be useful to us



Bernoulli Distribution

- Simple distribution that records probability of two discrete events
  - Coin flip

$$\mathrm{Bern}(x \mid \mu) = \mu^{x}(1-\mu)^{1-x}$$

- Parameter μ controls the distribution
  - It is the mean

Likelihood

- Given:
  - A sample of data
  - A probability distribution (a model) for the data
- We want to know how likely that data is given our model
  - P(D|model)
- Why do we want to do this?
  - When we learn, we want the model that best explains the data
    - Select a model that makes the data most probable

Likelihood

- Suppose we had a single coin flip (event)
- We can write the probability of this event

$$p(x \mid \mu) = \mu^{x}(1-\mu)^{1-x}$$

- What about the probability of n events?
- We need to assume independence
  - Each coin flip is independent
  - IID: independent and identically distributed
    - Each data point is obtained independently from the same distribution

Likelihood

- Writing out the entire likelihood

$$p(D \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- Remember our useful log trick

$$\log p(D \mid \mu) = \sum_{n=1}^{N} \log p(x_n \mid \mu) = \sum_{n=1}^{N} x_n \log \mu + (1-x_n)\log(1-\mu)$$

Logistic Regression: Conditional Log Likelihood

$$p(Y \mid X, w) = \prod_{i=1}^{n} p(y_i \mid x_i, w)$$

$$\ell(Y, X, w) = \log p(Y \mid X, w) = \sum_{i=1}^{n} \log p(y_i \mid x_i, w)$$

$$p(y = 1 \mid x, w) = \frac{1}{1 + e^{-w^T \cdot x}} \qquad p(y = 0 \mid x, w) = \frac{e^{-w^T \cdot x}}{1 + e^{-w^T \cdot x}}$$

Finding the Best Model

- How should we find the best model?
- First approach: maximum likelihood
  - Find the model parameters that maximize the likelihood of the data
- Function maximization
  - Take the derivative, set equal to 0

$$\mu = \frac{1}{N} \sum_{n=1}^{N} x_n$$

Maximum Likelihood

- Maximum likelihood estimation (MLE) is a common approach to estimating model parameters
- Straightforward: compute derivative, find parameters most likely to give observed data
- Problem: over fits data
  - Suppose we saw three coin flips, each one heads
  - MLE estimate $\mu = 1$
  - We saw similar behavior in linear regression
    - MLE for gaussians over-fits by under-estimating true variance

Bayesian Probabilities

- So far probabilities are frequencies of random events
- Bayesian view
  - Probabilities are quantifications of uncertainty
  - Not all events are repeatable
  - Ex. What grade will you get in this class
    - Over the semester our guess will get better
- Prior probability $p(a)$
- Likelihood of data given observations $p(D \mid a)$

Bayesian Estimation

- We have some idea of the value of μ
  - Most coins μ=0.5
- This is a prior
  - A belief as to the value of μ that isn't based on the data
  - Our prior: p(μ=0.5) is high
- We can use this prior to "smooth" our estimation

MAP

- Maximum posterior estimation (MAP)
  - Maximize the posterior distribution
- Posterior $\propto$ likelihood * prior
  - Likelihood: P(D|μ)
  - Prior: P(μ)
  - Posterior: P(μ|D)
- Find the most probable value of μ given the data (and the prior)

Building Classifiers

- Let's use probabilities to build a new classifier
- For regression we used likelihood
  - p(D|w) as a Gaussian
- For logistic regression we used conditional likelihood
  - We only had p(y|x)
  - Couldn't write out p(D|w)
- Let's try and write likelihood for classification
  - P(D|w)

Likelihood

- Each example is independent

$$p(Y,X) = \prod_{i=1}^{n} p(y_i, x_i)$$

- Rewrite probability as conditional

$$p(y_i, x_i) = p(y_i \mid x_i)\, p(x_i)$$

- Substitute

$$p(Y,X) = \prod_{i=1}^{n} p(y_i \mid x_i)\, p(x_i)$$

Likelihood

- Consider the first term $p(y_i \mid x_i)$
- How do we compute it?
- Let's use Bayes rule

$$p(y_i \mid x_i) = \frac{p(x_i \mid y_i)p(y_i)}{p(x_i)}$$

Likelihood

- Substitute $p(Y, X) = \prod_{i=1}^{n} \frac{p(x_i \mid y_i)p(y_i)}{p(x_i)} p(x_i)$
- Simplify $p(Y, X) = \prod_{i=1}^{n} p(x_i \mid y_i)p(y_i)$

Likelihood

- What is p(y)?
- The expected number of labels of each type
  - Easy to compute

Conditional

- What is p(x|y)?
  - Probability of generating example x given that it has label y
- How hard is this?
  - Remember that x is a vector
  - Equivalent to $p(x_{i1}, x_{i2}, x_{i3} \dots x_{iM} \mid y_i)$
  - Assuming binary features and binary label, how many parameters do we need?
    - 2 * (2^M-1) parameters!
      - (2^M-1) combinations for x
      - 2 labels

Assumptions

- Not enough data to observe all combinations even a single time
- Need to make simplifying assumptions

Conditional Independence

- **RV (random variable) X is conditionally independent of RV Y given RV Z if the probability of each is independent given Z**
- p(x,y|z) = p(x|z)p(y|z)
- Example
  - Probability that I need an umbrella and the ground is wet
  - Not independent! If its wet I probably need an umbrella because it is raining
  - I am told it is raining
  - Given this the probability that I need an umbrella is independent of the ground being wet
  - I gain no new information knowing that the ground is wet

Conditional Independence

- Assume each feature in x is independent given y
  - Once I know y each feature in x is independent
- Why is this helpful?

$$p(x_i \mid y_i) = \prod_{j=1}^{M} p(x_{ij} \mid y_i)$$

- This is a naïve assumption (it's very unlikely)

Conditional Independence

- How to estimate $p(x_{ij} \mid y_i)$?
  - Lots of data- every time feature x_{ij} occurs with y_i
- How many parameters do I need?
  - Before: 2 * (2^M-1)
  - Now: 2 * M
    - One parameter for each of M features
- Should be easier to learn so many fewer parameters

Maximum Likelihood Solution

- Solution is the mean of the probabilities in the data

$$p(y) = \frac{\text{number of times } y \text{ appears}}{\text{number of examples}}$$

$$p(x_j \mid y) = \frac{\text{number of times } x_j \text{ and } y \text{ appear together}}{\text{number of times } y \text{ appears}}$$

Predictions

- Given an example x, how do we make predictions?

$$\underset{y \in \{0,1\}}{\arg\max}\ p(y \mid x)$$

- Bayes rule and conditional assumption

$$\underset{y \in \{0,1\}}{\arg\max} \frac{\prod_{j=1}^{M} \{p(x_j \mid y)\} p(y)}{p(x)}$$

- Observe that p(x) does not depend on y

$$\underset{y \in \{0,1\}}{\arg\max} \prod_{j=1}^{M} \{p(x_j \mid y)\} p(y)$$

Naïve vs. Reality

- Positive: we now can parameterize our model
- Reality: naïve assumption very unlikely to be true
- Example:
  - Document classification: sports vs. finance
  - Each word in a document is a feature
  - Naïve assumption: once I know the topic is sports, every word is conditionally independent
    - Not true! Would be total nonsense
- Reality: works pretty well in practice
  - Assuming you don't have features that are highly correlated

Problem: Not enough data

- Recall that maximum likelihood is biased
  - Maximum likelihood over-fits the data
- In naïve Bayes this can be extreme
- What is the learned value of p(x|y) when I have never seen x and y together?
  - What happens to p(**x**|y)?

Solution: Bias Model Parameters

- Smoothing
  - Balances fitting the observed data with a prior bias
  - Laplacian smoothing (+1 smoothing)
    - Pretend we saw extra examples with every possible feature and label
    - Effectively add 1 to each count
- Bayesian methods
  - Put priors over model parameters

Naive Bayes

Fitting a function to data

- Fitting: Closed form solution: just count!
- Function: data likelihood (joint X and Y)
- Data: Naive feature conditional independence assumption