CS 2750: Machine Learning

Probability Review

Prof. Adriana Kovashka
University of Pittsburgh
February 29, 2016
Plan for today and next two classes

• Probability review
• Density estimation
• Naïve Bayes and Bayesian Belief Networks
Procedural View

• Training Stage:
  - Raw Data $\rightarrow x$ (Feature Extraction)
  - Training Data $\{(x,y)\} \rightarrow f$ (Learning)

• Testing Stage
  - Raw Data $\rightarrow x$ (Feature Extraction)
  - Test Data $x \rightarrow f(x)$ (Apply function, Evaluate error)
Statistical Estimation View

- Probabilities to rescue:
  - $x$ and $y$ are random variables
  - $D = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \sim P(X,Y)$

- IID: Independent Identically Distributed
  - Both training & testing data sampled IID from $P(X,Y)$
  - Learn on training set
  - Have some hope of generalizing to test set
Probability

• A is non-deterministic event
  – Can think of A as a boolean-valued variable

• Examples
  – A = your next patient has cancer
  – A = Rafael Nadal wins US Open 2016
Interpreting Probabilities

• What does $P(A)$ mean?

• Frequentist View
  – $\lim_{N \to \infty} \frac{\#(A \text{ is true})}{N}$
  – limiting frequency of a repeating non-deterministic event

• Bayesian View
  – $P(A)$ is your “belief” about $A$

• Market Design View
  – $P(A)$ tells you how much you would bet
Axioms of Probability Theory

- All probabilities between 0 and 1
  \[ 0 \leq P(A) \leq 1 \]

- True proposition has probability 1, false has probability 0.
  \[ P(\text{true}) = 1 \quad P(\text{false}) = 0 \]

- The probability of disjunction is:
  \[ P(A \lor B) = P(A) + P(B) - P(A \cap B) \]
Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{false}) = 0$
- $P(\text{true}) = 1$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

Event space of all possible worlds

Its area is 1

Worlds in which $A$ is true

$P(A) = \text{Area of reddish oval}$

Worlds in which $A$ is False
Interpreting the Axioms

- \(0 \leq P(A) \leq 1\)
- \(P(\text{false}) = 0\)
- \(P(\text{true}) = 1\)
- \(P(A \lor B) = P(A) + P(B) - P(A \land B)\)

The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true
Interpreting the Axioms

• 0 ≤ P(A) ≤ 1
• P(False) = 0
• P(True) = 1
• P(A v B) = P(A) + P(B) − P(A ^ B)

The area of A can't get any bigger than 1

And an area of 1 would mean all worlds will have A true
Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{false}) = 0$
- $P(\text{true}) = 1$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

Simple addition and subtraction
Joint Distribution

- The joint probability distribution for a set of random variables, \(X_1, \ldots, X_n\) gives the probability of every combination of values (an \(n\)-dimensional array with \(v^n\) values if all variables are discrete with \(v\) values, all \(v^n\) values must sum to 1): \(P(X_1, \ldots, X_n)\)

<table>
<thead>
<tr>
<th>Positive</th>
<th>Circle</th>
<th>Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>0.20</td>
<td>0.02</td>
</tr>
<tr>
<td>Blue</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative</th>
<th>Circle</th>
<th>Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>0.05</td>
<td>0.30</td>
</tr>
<tr>
<td>Blue</td>
<td>0.20</td>
<td>0.20</td>
</tr>
</tbody>
</table>

- The probability of all possible conjunctions (assignments of values to some subset of variables) can be calculated by summing the appropriate subset of values from the joint distribution.

\[
P(red \land circle) = \frac{P(red \land circle)}{P(red)}
\]

- Therefore, all conditional probabilities can also be calculated.

\[
P(positive \mid red \land circle)
\]
Marginal Distributions

\[ p(x, y) = \sum_{z \in \mathcal{Z}} p(x, y, z) \]

\[ p(x) = \sum_{y \in \mathcal{Y}} p(x, y) \]
Conditional Probabilities

- \( P(A \mid B) = \) In worlds where \( B \) is true, fraction where \( A \) is true

\[
P(A \mid B) = \frac{P(A \land B)}{P(B)}
\]

- Example
  - \( H \): “Have a headache”
  - \( F \): “Coming down with Flu”

\[
P(H) = 1/10
\]
\[
P(F) = 1/40
\]
\[
P(H \mid F) = 1/2
\]

Headaches are rare and flu is rarer, but if you're coming down with flu there's a 50-50 chance you'll have a headache.
Conditional Probabilities

• $P(Y=y \mid X=x)$

• What do you believe about $Y=y$, if I tell you $X=x$?

• $P($Rafael Nadal wins US Open 2016$)$?

• What if I tell you:
  – He has won the US Open twice
  – Novak Djokovic is ranked 1; just won Australian Open
Conditional Distributions

\[ p(x, y \mid Z = z) = \frac{p(x, y, z)}{p(z)} \]
Conditional Probabilities

$p(X,Y)$

$Y = 2$

$X$

$p(Y)$

$Y = 1$

$p(X)$

$X$

$p(X|Y = 1)$

$X$

Figures from Bishop
Chain rule

- Generalized product rule:

\[ P \left( \bigcap_{k=1}^{n} A_k \right) = \prod_{k=1}^{n} P \left( A_k \bigg| \bigcap_{j=1}^{k-1} A_j \right) \]

- Example:

\[ P(A_4, A_3, A_2, A_1) = P(A_4 \mid A_3, A_2, A_1) \cdot P(A_3 \mid A_2, A_1) \cdot P(A_2 \mid A_1) \cdot P(A_1) \]
Independence

• A and B are independent iff:

\[ P(A \mid B) = P(A) \]
\[ P(B \mid A) = P(B) \]

These two constraints are logically equivalent

• Therefore, if A and B are independent:

\[ P(A \mid B) = \frac{P(A \land B)}{P(B)} = P(A) \]
\[ P(A \land B) = P(A)P(B) \]
Independence

• Marginal: $P$ satisfies $(X \perp Y)$ if and only if
  
  $P(X=x, Y=y) = P(X=x) \ P(Y=y), \quad \forall x \in \text{Val}(X), \ y \in \text{Val}(Y)$

• Conditional: $P$ satisfies $(X \perp Y \mid Z)$ if and only if
  
  $P(X, Y|Z) = P(X|Z) \ P(Y|Z), \quad \forall x \in \text{Val}(X), \ y \in \text{Val}(Y), \ z \in \text{Val}(Z)$
Independent Random Variables

\[ P(x,y) \]

\[ X \perp Y \]

\[ p(x, y) = p(x)p(y) \]

for all \( x \in X, y \in Y \)
Other Concepts

• Expectation:

\[ \mathbb{E}[f] = \sum_x p(x) f(x) \]

• Variance:

\[ \text{var}[f] = \mathbb{E} \left[ (f(x) - \mathbb{E}[f(x)])^2 \right] \]

• Covariance:

\[ \text{cov}[x, y] = \mathbb{E}_{x,y} \left[ \{x - \mathbb{E}[x]\} \{y - \mathbb{E}[y]\} \right] \]

\[ = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \]
Entropy

- Measures the amount of ambiguity or uncertainty in a distribution:
  \[ H(p) = - \sum_x p(x) \log p(x) \]

- Expected value of \(-\log p(x)\) (a function which depends on \(p(x)\)).
- \(H(p) > 0\) unless only one possible outcome in which case \(H(p) = 0\).
- Maximal value when \(p\) is uniform.
- Tells you the expected "cost" if each event costs \(-\log p(\text{event})\)
KL-Divergence / Relative Entropy

- An asymmetric measure of the distance between two distributions:

\[ KL[p||q] = \sum_x p(x) \left( \log p(x) - \log q(x) \right) \]

- \( KL > 0 \) unless \( p = q \) then \( KL = 0 \)

- Tells you the extra cost if events were generated by \( p(x) \) but instead of charging under \( p(x) \) you charged under \( q(x) \).
Bayes Theorem

\[ P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \]

Simple proof from definition of conditional probability:

\[ P(H \mid E) = \frac{P(H \land E)}{P(E)} \quad \text{(Def. cond. prob.)} \]

\[ P(E \mid H) = \frac{P(H \land E)}{P(H)} \quad \text{(Def. cond. prob.)} \]

\[ P(H \land E) = P(E \mid H)P(H) = P(H \mid E)P(E) \]

QED: \[ P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \]

Adapted from Ray Mooney
Probabilistic Classification

- Let \( Y \) be the random variable for the class which takes values \( \{y_1,y_2,\ldots y_m\} \).
- Let \( X \) be the random variable describing an instance consisting of a vector of values for \( n \) features \( <X_1,X_2\ldots X_n> \), let \( x_k \) be a possible value for \( X \) and \( x_{ij} \) a possible value for \( X_i \).
- For classification, we need to compute \( P(Y=y_i \mid X=x_k) \) for \( i=1\ldots m \)
- However, given no other assumptions, this requires a table giving the probability of each category for each possible instance in the instance space, which is impossible to accurately estimate from a reasonably-sized training set.
  - Assuming \( Y \) and all \( X_i \) are binary, we need \( 2^n \) entries to specify \( P(Y=\text{pos} \mid X=x_k) \) for each of the \( 2^n \) possible \( x_k \)'s since \( P(Y=\text{neg} \mid X=x_k) = 1 - P(Y=\text{pos} \mid X=x_k) \)
  - Compared to \( 2^{n+1} - 1 \) entries for the joint distribution \( P(Y,X_1,X_2\ldots X_n) \)
Bayesian Categorization

- Determine category of $x_k$ by determining for each $y_i$

$$P(Y = y_i \mid X = x_k) = \frac{P(Y = y_i)P(X = x_k \mid Y = y_i)}{P(X = x_k)}$$

- $P(X=x_k)$ can be determined since categories are complete and disjoint.

$$P(X = x_k) = \sum_{i=1}^{m} P(Y = y_i)P(X = x_k \mid Y = y_i)$$

$$\sum_{i=1}^{m} P(Y = y_i \mid X = x_k) = \sum_{i=1}^{m} \frac{P(Y = y_i)P(X = x_k \mid Y = y_i)}{P(X = x_k)} = 1$$

Adapted from Ray Mooney
Bayesian Categorization (cont.)

- Need to know:
  - Priors: $P(Y=y_i)$
  - Conditionals (likelihood): $P(X=x_k \mid Y=y_i)$
- $P(Y=y_i)$ are easily estimated from data.
  - If $n_i$ of the examples in $D$ are in $y_i$ then $P(Y=y_i) = n_i / |D|$
- Too many possible instances (e.g. $2^n$ for binary features) to estimate all $P(X=x_k \mid Y=y_i)$.
- Need to make some sort of independence assumptions about the features to make learning tractable (more details later).

Adapted from Ray Mooney
Likelihood / Prior / Posterior

• A hypothesis is denoted as $h$; it is one member of the hypothesis space $H$
• A set of training examples is denoted as $D$, a collection of $(x, y)$ pairs for training
• $\Pr(h)$ – the *prior probability of the hypothesis* – without observing any training data, what’s the probability that $h$ is the target function we want?
• Pr(D) – the *prior probability of the observed data* – chance of getting the particular set of training examples D

• Pr(h|D) – the *posterior probability of h* – what is the probability that h is the target given that we’ve observed D?

• Pr(D|h) – the probability of getting D if h were true (a.k.a. *likelihood of the data*)

• Pr(h|D) = Pr(D|h)Pr(h)/Pr(D)
MAP vs MLE Estimation

• Maximum-a-posteriori (MAP) estimation:
  \[ h_{\text{MAP}} = \arg \max_h \Pr(h|D) \]
  \[ = \arg \max_h \Pr(D|h)\Pr(h)/\Pr(D) \]
  \[ = \arg \max_h \Pr(D|h)\Pr(h) \]

• Maximum likelihood estimation (MLE):
  \[ h_{\text{ML}} = \arg \max \Pr(D|h) \]