Probability Distributions on Structured Objects

September 17, 2013
Reminder

• HW1 is due at 11:59pm tonight

• There was some ambiguity in this assignment
• The TAs gave a lot of help, but in general, learning to work from incomplete specs is important
Probability Outline

• Why probability?
• Probability review
• Multinomials vs. exponential parameterization
• Locally vs. globally normalized models & partition functions
• Examples
Why Probability?

• Probability formalizes
  – The concept of models
  – The concept of data
  – The concept of learning
  – The concept of prediction (inference)

*Probability is expectation founded upon partial knowledge.*
Why Probability?

• What might we have partial knowledge about?
  – The state of the world (test data)
  – The reliability of our training data
  – The correctness of our model
  – The values of our parameters

\[ p(x \mid \text{partial knowledge}) \]
What is a Probability?

- **Limiting (relative) frequency of events**
  - in repeated (identical) experiments
- **Degree of belief**
  - Subjective conception
  - 40% chance of rain tomorrow in Pittsburgh
- **Viewpoint affects**
  - interpretation
  - **not** rules of probability calculus themselves
Discrete Distributions

Discrete distribution: $\Omega$ is *finite* or *countable*, but no bigger
Discrete Distributions

\[ \forall x \in \Omega, \quad f(x) \in [0, 1] \]

\[ \sum_{x \in \Omega} f(x) = 1 \]

An **event** is a subset (maybe one element) of the sample space, \( E \subseteq \Omega \)

\[ P(E) = \sum_{x \in E} f(x) \]
Random Variables

A random variable is a function from a random event from a set of possible outcomes $\Omega$ and a probability distribution $\rho$, a function from outcomes to probabilities.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$X(\omega) = \omega$$

$$\rho_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$
Random Variables

A random variable is a function from a random event from a set of possible outcomes \( \Omega \) and a probability distribution \( \rho \), a function from outcomes to probabilities.

\[
\Omega = \{1, 2, 3, 4, 5, 6\}
\]

\[
Y(\omega) = \begin{cases} 
0 & \text{if } \omega \in \{2, 4, 6\} \\
1 & \text{otherwise}
\end{cases}
\]

\[
\rho_Y(y) = \begin{cases} 
\frac{1}{2} & \text{if } y = 0, 1 \\
0 & \text{otherwise}
\end{cases}
\]
Sampling Notation

\[ x = 4 \times z + 1.7 \]
Sampling Notation

\[ x = 4 \times z + 1.7 \]

\[ y \sim \text{Distribution}(\theta) \]
Sampling Notation

\[ x = 4 \times z + 1.7 \]

\[ y \sim \text{Distribution}(\theta) \]

\[ y' = y \times x \]

Random variable
Joint Probability

- Probability over multiple event types
- Tool for reasoning about dependent (correlated) events

A **joint probability distribution** is a probability distribution over r.v.'s with the following form:

\[
Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix}
\]

\[
\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) = 1 \quad \rho_Z \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) \geq 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}
\]
Joint Probability

- Probability over multiple event types
- Tool for reasoning about dependent (correlated) events

A joint probability distribution is a probability distribution over r.v.’s with the following form:

\[ Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} \]

\[ \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) = 1 \quad \rho_Z \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) \geq 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} \]
Joint Probability

- Probability over multiple event types
- Tool for reasoning about dependent (correlated) events

A joint probability distribution is a probability distribution over r.v.’s with the following form:

$$Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix}$$

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) = 1$$

$$\rho_Z \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) \geq 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$
Joint Probability

- Probability over multiple event types
- Tool for reasoning about dependent (correlated) events

A **joint probability distribution** is a probability distribution over r.v.’s with the following form:

$$Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix}$$

For any $x \in \mathcal{X}, y \in \mathcal{Y}$:

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z \begin{bmatrix} x \\ y \end{bmatrix} = 1$$

$$\rho_Z \begin{bmatrix} x \\ y \end{bmatrix} \geq 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

DNA sequence

Proteins
\[ \Omega = \{1, 2, 3, 4, 5, 6\} \]

\[ X(\omega) = \omega \]
\[ \Omega = \{1, 2, 3, 4, 5, 6\} \]

\[ X(\omega) = \omega \]

\[ \Omega = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \]
\[ (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \]
\[ (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \]
\[ (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \]
\[ (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \]
\[ (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6), \} \]

\[ X(\omega) = \omega_1 \quad Y(\omega) = \omega_2 \]

\[ \rho_{X,Y}(x, y) = \begin{cases} \frac{1}{36} & \text{if } (x, y) \in \Omega \\ 0 & \text{otherwise} \end{cases} \]
\[ \Omega = \{1, 2, 3, 4, 5, 6\} \]

\[ X(\omega) = \omega \]

\[ \Omega = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\
(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\
(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\
(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\
(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\
(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6), \} \]

\[ X(\omega) = \omega_1 \quad Y(\omega) = \omega_2 \]

\[ \rho_{X,Y}(x, y) = \begin{cases} \frac{x+y}{252} & \text{if } (x, y) \in \Omega \\ 0 & \text{otherwise} \end{cases} \]
Marginal Probability

\[ p(X = x, Y = y) = \rho_{X,Y}(x, y) \]

\[ p(X = x) = \sum_{y' \in Y} p(X = x, Y = y') \]

\[ p(Y = y) = \sum_{x' \in X} p(X = x', Y = y) \]

\[ \Omega = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\} \]

\[ p(X = 4) = \sum_{y' \in [1,6]} p(X = 4, Y = y') \]

\[ p(Y = 3) = \sum_{x' \in [1,6]} p(X = x', Y = 3) \]
Marginal Probability

Sample space

- (NN, cat)
- (NN, sloth)
- (NN, book)
- (JJ, fuzzy)
- (VB, book)
- (RB, quickly)
Marginal Probability

Sample space

\[ p(t = \text{NN}) \]

- (NN, ·)
- (JJ, fuzzy)
- (VB, book)
- (RB, quickly)

\[ (\text{NN}, \cdot) \]
Marginal Probability

Sample space

\[ p(w = \text{book}) \]
Marginal Probability

Sample space

- (NN, cat)
- (NN, sloth)
- (NN, book)
- (JJ, fuzzy)
- (VB, book)
- (RB, quickly)
Marginal Probabilities

• In a joint model of word and tag sequences $p(w,t)$
  – The probability of a word sequence $p(w)$
  – The probability of a tag sequence $p(t)$
  – The probability of a word sequence with the word “cat” somewhere in it
  – The probability of a tag sequence containing three verbs in a row
Conditional Probability

The **conditional probability** is defined as follows:

\[
p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{\text{joint probability}}{\text{marginal}}
\]

This assumes \( p(Y = y) \neq 0 \)

We can construct joint probability distributions out of conditional distributions:

\[
p(x \mid y)p(y) = p(x, y) = p(y \mid x)p(x)
\]
Conditional Probability Distributions

The **conditional probability distribution** of a variable $X$ given a variable $Y$ has the following properties:

\[ \forall y \in Y, \sum_{x \in X} p(X = x \mid Y = y) = 1 \]
Conditional Probability

Sample space

- (NN, cat)
- (NN, sloth)
- (NN, book)
- (JJ, fuzzy)
- (VB, book)
- (RB, quickly)
Conditional Probability

Sample space

- (NN, cat)
- (NN, sloth)
- (NN, book)
- (JJ, fuzzy)
- (VB, book)
- (RB, quickly)

\[ p(\cdot | w = \text{book}) \]
Conditional Probabilities

• In a joint model of word and tag sequences $p(w,t)$
  – The probability of a tag sequence given a word sequence $p(t \mid w)$
  – The probability of a word sequence given a tag sequence $p(w \mid t)$
Joint and Marginal Probabilities

• In a joint model of word and tag sequences $p(w, t)$
  
  – The probability that the 3rd tag is VERB, given $w = “\text{Time flies like an arrow}”$
    $p(t_3 = \text{VERB} \mid w = \text{Time flies like an arrow})$

  – The probability that the 3rd word is like, given $w = “\text{Time flies _____ an arrow}”$, $t_3 = \text{VERB}$
    $p(t_3 = \text{like} \mid w = \text{Time flies _____ an arrow, } t_3 = \text{VERB})$
Chain Rule

\[ p(a, b, c, d, \ldots) = p(a) \times p(b \mid a) \times p(c \mid a, b) \times p(d \mid a, b, c) \times \cdots \]
Bayes Rule

\[ p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} \]

Posterior \quad Likelihood \quad Prior

Evidence

\[ = \frac{p(y \mid x)p(x)}{\sum_{x'} p(y \mid x')p(x')} \]
Independence

Two r.v.’s are **independent** iff

\[ p(X = x, Y = y) = p(X = x) \times p(Y = y) \]

Equivalently (prove with def. of cond. prob.)

\[ p(X = x \mid Y = y) = p(X = x) \]

Alternatively,

\[ p(Y = y \mid X = x) = p(Y = y) \]
Conditional Independence

Two equivalent statements of conditional independence:
\[ p(a, c \mid b) = p(a \mid b)p(c \mid b) \]
and:
\[ p(a \mid b, c) = p(a \mid b) \]

“If I know B, then C doesn’t tell me about A”
\[ p(a \mid b, c) = p(a \mid b) \]
\[ p(a, b, c) = p(a \mid b, c)p(b, c) \]
\[ = p(a \mid b, c)p(b \mid c)p(c) \]
Conditional Independence

Two equivalent statements of conditional independence:

\[ p(a, c \mid b) = p(a \mid b)p(c \mid b) \]

and:

\[ p(a \mid b, c) = p(a \mid b) \]

“If I know B, then C doesn’t tell me about A”

\[ p(a \mid b, c) = p(a \mid b) \]

\[ p(a, b, c) = p(a \mid b, c)p(b, c) \]

\[ = p(a \mid b, c)p(b \mid c)p(c) \]

\[ = p(a \mid b)p(b \mid c)p(c) \]
Conditional Independence

• Useful thing to assume when designing models
  – Limit the variables that influence distributions
  – Classical example: Markov assumption

• Questions
  – Does conditional independence imply marginal independence?
  – Does marginal independence imply conditional independence?
Expected Values

\[ \mathbb{E}_{p(X=x)} [f(x)] = \sum_{x \in \mathcal{X}} p(X = x) \times f(x) \]

Some special expectations:

\[ p(X = y) = \mathbb{E}_{p(X=x)} [\mathbb{I}_{x=y}] \]

\[ H(X) = \mathbb{E}_{p(X=x)} [- \log_2 x] \]
Categorical (Multinomial) Distributions

- Generalized model of a distribution to \( k \) dimensions
- Option 1: Parameters lie on the **\( k \)-simplex**

\[
\Delta^k = \left\{ (\theta_1, \theta_2, \ldots, \theta_k) \mid \sum_{i=1}^{k} \theta_i = 1 \land \theta_i \geq 0 \ \forall \ i \in [0, k] \right\}
\]
Log-linear Parameterization

\[ p(x) = \frac{\exp \mathbf{w}^\top f(x)}{Z} \]

where \( Z = \sum_{x' \in X} \exp \mathbf{w}^\top f(x) \)

Assumption: \( Z \) converges
Categorical (Multinomial) Distributions

• “Naïve” parameterization
  – k outcomes, k(-1) independent parameters
  – Model as tables of (conditional) probabilities
  – MLE estimation (given fully observed data) is easy

• Log-linear parameterization
  – k outcomes, n, possibly overlapping parameters
    • Share statistical strength across “related” events
    • How are elements related? Depends how you define f
Locally Normalized Models

• Structure as the result of a **discrete time branching process**
  – Start in a known initial state, carry out stochastic steps (parameterized using multinomials) until some termination condition is met
  – Steps are (conditionally) independent of one another: probabilities multiply
  – *Total probability is the probability of the steps*

• Usually for joint (generative) models
  – not always though (see Appendix D.2)
S

1.0 \times p(NP \; VP \mid S)
1.0 \times p(\text{NP VP | S})
\times p(\text{JJ NN | NP})
\[ 1.0 \times p(\text{NP VP} \mid S) \times p(\text{JJ NN} \mid \text{NP}) \times p(\text{V} \mid \text{VP}) \]
1.0 x p(NP VP | S) 
  x p(JJ NN | NP) 
  x p(V | VP) 
  x p(\textit{angry} | JJ)
angry dogs

1.0 x p(NP VP | S)
x p(JJ NN | NP)
x p(V | VP)
x p(\textit{angry} | JJ)
x p(\textit{dogs} | NN)
\[ p(\tau, x) = \prod_{r \in G} p(r | G)^{f(r \in \tau)} \]
1.0 \times p(\text{NP VP} \mid S) \\
\times p(\text{JJ NN} \mid \text{NP}) \\
\times p(\text{V} \mid \text{VP}) \\
\times p(\text{angry} \mid \text{JJ}) \\
\times p(\text{dogs} \mid \text{NN}) \\
\times p(\text{bark} \mid \text{V})
Here’s an alternative way of building a tree and string:

S 1.0
Here’s an alternative way of building a tree and string:

$S$

$1.0 \times p(2 \text{ kids} \mid S)$
Here’s an alternative way of building a tree and string:

\[ 1.0 \times p(2 \text{ kids} \mid S) \times p(NP \mid S, n=1, \text{total}=2) \]
Here’s an alternative way of building a tree and string:

1.0 \times p(2 \text{ kids} \mid S) \\
\times p(NP \mid S, n=1, \text{total}=2) \\
\times p(VP \mid S, n=2, \text{total}=2)
Here’s an alternative way of building a tree and string:

1.0 x p(2 kids | S)
 x p(NP | S, n=1, total=2)
 x p(VP | S, n=2, total=2)
 x p(1 kid | VP)
Here’s an alternative way of building a tree and string:

\[
\begin{align*}
S & \quad 1.0 \times p(2 \text{ kids } | \ S) \\
& \quad \times p(\text{NP } | \ S, n=1, \text{ total}=2) \\
& \quad \times p(\text{VP } | \ S, n=2, \text{ total}=2) \\
& \quad \times p(1 \text{ kid } | \ VP)
\end{align*}
\]
Here's an alternative way of building a tree and string:
Here’s an alternative way of building a tree and string:

```
1.0 x p(2 kids | S)
x p(NP | S, n=1, total=2)
x p(VP | S, n=2, total=2)
x p(1 kid | VP)
x p(1 kid | VP, S)
```
Choosing a Model

• Independence is a property of distributions
  – Look at distributions in the wild, figure out what independence assumptions hold

• Dependence makes modeling more expensive
  – How big does your CKY chart have to be if you have “grandparent” annotation?
Parameterization

- For each step in the branching process
  - We have a multinomial distribution
  - We can use independent parameters (on simplex)
  - We can use log-linear models
    - “Locally normalized model” (cf. Appendix D.2)
    - Z is “local” to the decision being made
Globally Normalized Models

• Extension of the exponential parameterization to structured output spaces

\[ p(x) = \frac{\exp w^\top F(x)}{Z} \]

where \( Z = \sum_{x' \in \mathcal{X}} \exp w^\top F(x') \)
Conditional Random Fields

\[ p(y \mid x) = \frac{\exp w^\top F(x)}{Z(x)} \]

\[ Z(x) = \sum_{y' \in \mathcal{Y}_x} \exp w^\top F(x) \]
Conditional Random Fields

\[ p(y \mid x) = \frac{\exp w^\top F(x, y)}{Z(x)} \]

\[ Z(x) = \sum_{y' \in \mathcal{Y}_x} \exp w^\top F(x, y') \]

Decoding is nice:

\[ y^* = \arg \max_{y \in \mathcal{Y}_x} \frac{\exp w^\top F(x, y)}{Z(x)} \]

\[ = \arg \max_{y \in \mathcal{Y}_x} \exp w^\top F(x, y) \]

\[ = \arg \max_{y \in \mathcal{Y}_x} w^\top F(x, y) \]
Conditional Random Fields

\[ F(x, y) = \sum_{C \in G} f(C) \]
Comparison of Feature-Based Models

• Locally Normalized Models
  – Good joint models
  – Easy to training
  – Downside: decoding can be expensive

• Globally Normalized Models
  – Very popular conditional models (CRFs)
  – Challenge: computing $Z$ / training
  – Advantage: decoding can be cheap