Introduction to Probability for Graphical Models

CSC 412
Kaustav Kundu
Thursday January 14, 2016

*Most slides based on Kevin Swersky’s slides, Inmar Givoni’s slides, Danny Tarlow’s slides, Jasper Snoek’s slides, Sam Roweis ‘s review of probability, Bishop’s book, and some images from Wikipedia
Outline

• Basics
• Probability rules
• Exponential family models
• Maximum likelihood
• Conjugate Bayesian inference (time permitting)
Why Represent Uncertainty?

• The world is full of uncertainty
  – “What will the weather be like today?”
  – “Will I like this movie?”
  – “Is there a person in this image?”

• We’re trying to build systems that understand and (possibly) interact with the real world

• We often can’t prove something is true, but we can still ask how likely different outcomes are or ask for the most likely explanation

• Sometimes probability gives a concise description of an otherwise complex phenomenon.
Why Use Probability to Represent Uncertainty?

• Write down simple, reasonable criteria that you'd want from a system of uncertainty (common sense stuff), and you always get probability.

• Cox Axioms (Cox 1946); See Bishop, Section 1.2.3

• We will restrict ourselves to a relatively informal discussion of probability theory.
Notation

- A random variable $X$ represents outcomes or states of the world.
- We will write $p(x)$ to mean $\text{Probability}(X = x)$
- Sample space: the space of all possible outcomes (may be discrete, continuous, or mixed)
- $p(x)$ is the probability mass (density) function
  - Assigns a number to each point in sample space
  - Non-negative, sums (integrates) to 1
  - Intuitively: how often does $x$ occur, how much do we believe in $x$. 
Joint Probability Distribution

- $\text{Prob}(X=x, Y=y)$
  - “Probability of $X=x$ and $Y=y$”
  - $p(x, y)$

Conditional Probability Distribution

- $\text{Prob}(X=x \mid Y=y)$
  - “Probability of $X=x$ given $Y=y$”
  - $p(x \mid y) = p(x, y) / p(y)$
The Rules of Probability

- **Sum Rule** (marginalization/summing out):
  \[ p(x) = \sum_y p(x, y) \]

  \[ p(x_1) = \sum_{x_2} \sum_{x_3} \ldots \sum_{x_N} p(x_1, x_2, \ldots, x_N) \]

- **Product/Chain Rule**:
  \[ p(x, y) = p(y | x)p(x) \]
  \[ p(x_1, \ldots, x_N) = p(x_1)p(x_2 | x_1)\ldots p(x_N | x_1, \ldots, x_{N-1}) \]
Bayes’ Rule

- One of the most important formulas in probability theory

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

- This gives us a way of “reversing” conditional probabilities
Independence

• Two random variables are said to be independent iff their joint distribution factors

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

• Two random variables are conditionally independent given a third if they are independent after conditioning on the third

\[ p(x, y \mid z) = p(y \mid x, z)p(x \mid z) = p(y \mid z)p(x \mid z) \quad \forall z \]
Continuous Random Variables

• Outcomes are real values. Probability density functions define distributions.
  – E.g.,

\[ P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \]

• Continuous joint distributions: replace sums with integrals, and everything holds
  – E.g., Marginalization and conditional probability

\[ P(x, z) = \int P(x, y, z) = \int P(x, z \mid y) P(y) \]
Summarizing Probability Distributions

• It is often useful to give summaries of distributions without defining the whole distribution (E.g., mean and variance)

• Mean:
  \[ E[x] = \langle x \rangle = \int x \cdot p(x) \, dx \]

• Variance:
  \[ \text{var}(x) = \int (x - E[x])^2 \cdot p(x) \, dx \]
  \[ = E[x^2] - E[x]^2 \]
Summarizing Probability Distributions

- It is often useful to give summaries of distributions without defining the whole distribution (E.g., mean and variance)

- **Mean:**
  \[ E[x] = \langle x \rangle = \int x \cdot p(x) \, dx \]

- **Variance:**
  \[ \text{var}(x) = \int (x - E[x])^2 \cdot p(x) \, dx \]
  \[ = E[x^2] - (E[x])^2 \]
Exponential Family

- Family of probability distributions
- Many of the standard distributions belong to this family
  - Bernoulli, Binomial/Multinomial, Poisson, Normal (Gaussian), Beta/Dirichlet,…
- Share many important properties
  - e.g. They have a conjugate prior (we’ll get to that later. Important for Bayesian statistics)
Definition

- The exponential family of distributions over $x$, given parameter $\eta$ (eta) is the set of distributions of the form

$$p(x \mid \eta) = h(x) g(\eta) \exp \{\eta^T u(x)\}$$

- $x$ - scalar/vector, discrete/continuous
- $\eta$ - ‘natural parameters’
- $u(x)$ - some function of $x$ (sufficient statistic)
- $g(\eta)$ - normalizer

$$g(\eta) \int h(x) \exp \{\eta^T u(x)\} dx = 1$$

- $h(x)$ - base measure (often constant)
Sufficient Statistics

• Vague definition: called so because they completely summarize a distribution.
• Less vague: they are the only part of the distribution that interacts with the parameters and are therefore sufficient to estimate the parameters.
Example 1: Bernoulli

- Binary random variable - $X \in \{0,1\}$
- $p(\text{heads}) = \mu$  $\mu \in [0,1]$
- Coin toss

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

\[ p(x \mid \eta) = h(x)g(\eta) \exp \{\eta^T u(x)\} \]

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

\[ = \exp \{x \ln \mu + (1 - x) \ln(1 - \mu)\} \]

\[ = (1 - \mu) \exp \{\ln\left(\frac{\mu}{1 - u}\right)x\} \]

\[ p(x \mid \eta) = \sigma(-\eta) \exp(\eta x) \]

\[ h(x) = 1 \]

\[ u(x) = x \]

\[ \eta = \ln\left(\frac{\mu}{1 - \mu}\right) \Rightarrow \mu = \sigma(\eta) = \frac{1}{1 + e^{-\eta}} \]

\[ g(\eta) = \sigma(-\eta) \]
Example 2: Multinomial

- \( p(\text{value } k) = \mu_k \)
  \[ \mu_k \in [0,1], \sum_{k=1}^{M} \mu_k = 1 \]

- For a single observation - die toss
  - Sometimes called Categorical

- For multiple observations
  - integer counts on \( N \) trials
  - \( \text{Prob}(1 \text{ came out } 3 \text{ times, } 2 \text{ came out once, ..., 6 came out } 7 \text{ times if I tossed a die } 20 \text{ times}) \)

\[
P(x_1, \ldots, x_M \mid \mu) = \frac{N!}{\prod_k x_k!} \prod_{k=1}^{M} \mu_k^{x_k}
\]
Example 2: Multinomial (1 observation)

\[ p(x \mid \eta) = h(x) g(\eta) \exp \{ \eta^T u(x) \} \]

\[ P(x_1, \ldots, x_M \mid \mu) = \prod_{k=1}^{M} \mu_k^{x_k} \]

\[ = \exp \left\{ \sum_{k=1}^{M} x_k \ln \mu_k \right\} \]

\[ p(x \mid \eta) = \exp(\eta^T x) \]

\[ h(x) = 1 \]

\[ u(x) = x \]

Parameters are not independent due to constraint of summing to 1, there’s a slightly more involved notation to address that, see Bishop 2.4
Example 3: Normal (Gaussian) Distribution

- Gaussian (Normal)

\[ p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \]
Example 3: Normal (Gaussian) Distribution

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

- \( \mu \) is the mean
- \( \sigma^2 \) is the variance
- Can verify these by computing integrals.

E.g.,

\[
\int_{x \to \infty} x \cdot \frac{1}{\sqrt{2\pi \sigma}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} dx = \mu
\]
Example 3: Normal (Gaussian) Distribution

- **Multivariate Gaussian**

\[ P(x \mid \mu, \Sigma) = |2\pi \Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]
Example 3: Normal (Gaussian) Distribution

• Multivariate Gaussian

\[ p(x \mid \mu, \Sigma) = \left| 2\pi \Sigma \right|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]

• \( x \) is now a vector
• \( \mu \) is the mean vector
• \( \Sigma \) is the covariance matrix
Important Properties of Gaussians

• All marginals of a Gaussian are again Gaussian
• Any conditional of a Gaussian is Gaussian
• The product of two Gaussians is again Gaussian
• Even the sum of two independent Gaussian RVs is a Gaussian.
• Beyond the scope of this tutorial, but very important: marginalization and conditioning rules for multivariate Gaussians.
Gaussian marginalization visualization
Exponential Family Representation

\[ p(x \mid \eta) = h(x)g(\eta) \exp \{ \eta^T u(x) \} \]

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

\[
= \frac{1}{\sqrt{2\pi \sigma}} \exp \left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x + \frac{1}{2\sigma^2} \mu^2 \right\} = \\
(2\pi)^{-\frac{1}{2}} (-2\eta_2)^{\frac{1}{2}} \exp \left( \frac{\eta_1^2}{4\eta_2} \right) \exp \left\{ \frac{\mu}{\sigma^2} \frac{1}{2\sigma^2} \eta_2 \right\} \left[ \begin{array}{c} \mu \\ \eta^T \\ u(x) \end{array} \right]
\]
Example: Maximum Likelihood For a 1D Gaussian

- Suppose we are given a data set of samples of a Gaussian random variable $X$, $D=\{x^1, \ldots, x^N\}$ and told that the variance of the data is $\sigma^2$.

*Need to assume data is independent and identically distributed (i.i.d.)*
Example: Maximum Likelihood For a 1D Gaussian

What is our best guess of $\mu$?

- We can write down the likelihood function:

$$p(d \mid \mu) = \prod_{i=1}^{N} p(x^i \mid \mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2 \pi \sigma}} \exp \left\{ - \frac{1}{2 \sigma^2} (x^i - \mu)^2 \right\}$$

- We want to choose the $\mu$ that maximizes this expression
  - Take log, then basic calculus: differentiate w.r.t. $\mu$, set derivative to 0, solve for $\mu$ to get sample mean

$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$$
Example: Maximum Likelihood For a 1D Gaussian

Maximum Likelihood
ML estimation of model parameters for Exponential Family

\[ p(D | \eta) = p(x_1, ..., x_N) = \left( \prod h(x_n) \right) g(\eta)^N \exp \{ \eta^T \sum_n u(x_n) \} \]

\[ \partial \frac{\ln(p(D | \eta))}{\partial \eta} = ..., \text{set to 0, solve for } \nabla g(\eta) \]

\[ -\nabla \ln g(\eta_{ML}) = \frac{1}{N} \sum_{n=1}^{N} u(x_n) \]

• Can in principle be solved to get estimate for eta.
• The solution for the ML estimator depends on the data only through sum
  over u, which is therefore called **sufficient statistic**
• What we need to store in order to estimate parameters.
Bayesian Probabilities

\[ p(\theta \mid d) = \frac{p(d \mid \theta)p(\theta)}{p(d)} \]

- \( p(d \mid \theta) \) is the likelihood function
- \( p(\theta) \) is the prior probability of (or our prior belief over) \( \theta \)
  - our beliefs over what models are likely or not before seeing any data
- \( p(d) = \int p(d \mid \theta)P(\theta)d\theta \) is the normalization constant or partition function
- \( p(\theta \mid d) \) is the posterior distribution
  - Readjustment of our prior beliefs in the face of data
Example: Bayesian Inference For a 1D Gaussian

• Suppose we have a prior belief that the mean of some random variable $X$ is $\mu_0$ and the variance of our belief is $\sigma_0^2$

• We are then given a data set of samples of $X$, $d=\{x^1, \ldots, x^N\}$ and somehow know that the variance of the data is $\sigma^2$

What is the posterior distribution over (our belief about the value of) $\mu$?
Example: Bayesian Inference For a 1D Gaussian
Example: Bayesian Inference For a 1D Gaussian

Prior belief

\[
x^1 \ x^2 \ \ldots \ \mu_0 \ x^N
\]
Example: Bayesian Inference For a 1D Gaussian

• Remember from earlier
  
  \[ p(\mu \mid d) = \frac{p(d \mid \mu) p(\mu)}{p(d)} \]

• \( p(d \mid \mu) \) is the likelihood function

\[
p(d \mid \mu) = \prod_{i=1}^{N} P(x^i \mid \mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi \sigma}} \exp\left\{ -\frac{1}{2\sigma^2} (x^i - \mu)^2 \right\}
\]

• \( p(\mu) \) is the prior probability of (or our prior belief over) \( \mu \)

\[
p(\mu \mid \mu_0, \sigma_0) = \frac{1}{\sqrt{2\pi \sigma_0}} \exp\left\{ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right\}
\]
Example: Bayesian Inference For a 1D Gaussian

\[ p(\mu \mid D) \propto p(D \mid \mu) p(\mu) \]

\[ p(\mu \mid D) = \text{Normal}(\mu \mid \mu_N, \sigma_N) \]

where \[ \mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML} \]

\[ \frac{1}{\sigma^2_N} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \]
Example: Bayesian Inference For a 1D Gaussian

Prior belief
Example: Bayesian Inference For a 1D Gaussian

Prior belief

Maximum Likelihood
Example: Bayesian Inference For a 1D Gaussian

Prior belief

Maximum Likelihood

Posterior Distribution
Conjugate Priors

- Notice in the Gaussian parameter estimation example that the functional form of the posterior was that of the prior (Gaussian).
- Priors that lead to that form are called ‘conjugate priors’.
- For any member of the exponential family there exists a conjugate prior that can be written like

\[ p(\eta | \chi, \nu) = f(\chi, \nu) g(\eta)^\nu \exp\{\nu \eta^T \chi\} \]

- Multiply by likelihood to obtain posterior (up to normalization) of the form

\[ p(\eta | D, \chi, \nu) \propto g(\eta)^{\nu+N} \exp\{\eta^T (\sum_{n=1}^{N} u(x_n) + \nu \chi)\} \]

- Notice the addition to the sufficient statistic.
- \( \nu \) is the effective number of pseudo-observations.
Conjugate Priors - Examples

• Beta for Bernoulli/binomial
• Dirichlet for categorical/multinomial
• Normal for Normal