The Big Picture

Summary of course so far

- compact representations of high-dimensional distributions
  - Bayes nets, MRFs, CRFs
  - conditional independence, graph structure, factorization
- inference
  - conditioning, marginalization
  - variable elimination, message passing
- learning
  - Bayes nets: counting
  - MRFs/CRFs: numerical optimization of log-likelihood, inference is key subroutine

What's left?

- Inference (and therefore learning) not tractable for many models
  - approximate inference
  - variational methods
- Other types of probability distributions (continuous, parametric, ...)

Based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)
Today

- A bit of probability: continuous distributions, expectations
- Exponential families: very general class of distributions
  - includes MRFs
  - "redo" learning in much more general way

Continuous Random Variables and Density Functions

How to define the distribution of a random variable $X \in \mathbb{R}^d$?
The random variable $X \in \Omega$ has **density function** $p : \Omega \rightarrow \mathbb{R}^+$ if

$$P(X \in A) = \int_A p(x) dx$$

$1 = P(X \in \Omega) = \int_{\Omega} p(x) dx$

Implies $p(x) \geq 0$, $\int_{\Omega} p(x) = 1$.

**Note**: a pmf is a density function (integral over finite set $\equiv$ sum)

Example: Normal Distribution

The univariate normal (or Gaussian) distribution is the most well known continuous
distribution. It has density

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

$\mu \in \mathbb{R}$: location, mean, mode
$\sigma^2 \geq 0$: spread, scale, variance

$P(a \leq X \leq b)$

$= \int_a^b p(x; \mu, \sigma^2) dx$
How to Think About a Density

A density is “like” a probability. For $X \in \mathbb{R}$ with density $p(x)$

$$P(X \in [x, x + \epsilon]) = \int_x^{x+\epsilon} p(x)dx \approx \epsilon p(x)$$

$$p(x) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} P(X \in [x, x + \epsilon])$$

The density can be thought of as the probability of $X$ landing in a tiny interval around $x$ (divided the width of the interval).

The standard rules of probability (conditioning, marginalization) usually translate to densities in a straightforward way.

Example: Multivariate Normal Distribution

A multivariate normal (or Gaussian) random variable $X \in \mathbb{R}^n$ has density

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

- $\mu \in \mathbb{R}^n$: mean, mode
- $\Sigma \in \mathbb{R}^{n \times n}$: covariance matrix, defines scale and orientation
  - Must be positive definite (PSD): $x^T \Sigma x > 0$ for all $x \in \mathbb{R}^n$. (Equivalently, all eigenvalues positive.)

Visualization

Sequence of examples due to Andrew Ng / Stanford

Multivariate Gaussian

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$
Examples: Symmetric

\[ \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = 0.6I; \quad \Sigma = 2I. \]

Contours

\[ \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}. \]

Examples: Non-Symmetric

\[ \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}. \]

Mean

- Change mu: move mean of density around

\[ \mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \quad \mu = \begin{bmatrix} -0.5 \\ 0 \end{bmatrix}; \quad \mu = \begin{bmatrix} -1 \\ -1.5 \end{bmatrix}. \]
Marginal and Conditional Densities

- Definitions from pmfs usually translate to densities
- Suppose $p(x, y)$ is a density for $(X, Y)$. The marginal and conditional densities are

$$p(y) = \int p(x, y) dx$$

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

Expectations

Given a random variable $X$ with pmf or density $p(x)$ and a function $f(X)$, the expected value $\mathbb{E}[f(X)]$ is

$$\mathbb{E}[f(X)] = \sum_x p(x) f(x) \quad \text{discrete}$$

$$\mathbb{E}[f(X)] = \int p(x) f(x) dx \quad \text{continuous}$$

The sum/integral is over all possible values of $x$.

We often write this as $\mathbb{E}_{p(x)}[f(X)]$ to make the distribution clear.

Mean and Variance

The moments of a distribution are expectations of polynomials, e.g. $f(x) = (x - c)^d$ for scalars.

The mean is

$$\mu = \mathbb{E}[X] = \int p(x) x dx$$

Let $\mu = \mathbb{E}[X]$. The variance is

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2]$$

$$\text{Var}(X) = \mathbb{E}[(X - \mu)(X - \mu)^T]$$

$X$ scalar

$X$ vector

$\sum_i c_i(x_i, x_j)$
Marginal and conditional means use marginal and conditional densities:

\[ E_{p(x,y)}[y] = E_{p(y)}[y] \quad \text{marginal} \]

\[ E_{p(x,y)}[x|y = y] = E_{p(x|y)}[x] \quad \text{conditional} \]

In the vector case, \( \text{Var}(X) \) is the covariance matrix.

### Linearity of Expectation

For \( X, a, b \in \mathbb{R} \):

\[ E[aX + b] = aE[X] + b \]

For vectors \( X \) and \( b \) and matrix \( A \):

\[ E[AX + b] = A E[X] + b \]

**Proof**: write out expectation, use linearity of sum/integral

### Variance is Positive (Semi-Definite)

A covariance matrix \( \text{Var}(X) \) is always positive semi-definite.

**Proof** (scalar): \( E[(X - \mu)^2] \geq 0 \) because the integrand is non-negative

**Proof** (vector): let \( z \) be any vector and \( \mu = E[X] \). Then

\[
E[(X - \mu)^T (X - \mu) z] = E[(X - \mu)^T (X - \mu)'] E[z']
= E[E[(X - \mu)^T (X - \mu)'] E[z']]
= E[0]
= 0
\]

### Significance

Expectations are important! But, like many important things, they can be hard to compute:

**Example**: suppose \( p(x) \) is an MRF, then

\[
P(X_u = a, X_v = b) = E_{p(x)}[1[X_u = a, X_v = b]]
\]

Inference = computing expectations = hard in general

We will come back to approximating expectations and approximate inference
Exponential Families

An exponential family defines a set of distributions with densities of the form

\[ p_\theta(x) = h(x) \exp(\theta^T T(x) - A(\theta)) \]

- \( \theta \): “natural parameters”
- \( T(x) \): “sufficient statistics”
- \( A(\theta) \): “log-partition function”
- \( h(x) \): “base measure” (we’ll usually ignore)

Interpretation \((h(x) = 1)\)

\[ p_\theta(x) = \exp(\theta^T T(x) - A(\theta)) \cdot \exp(\theta^T T(x) \cdot \exp(-A(\theta))) \]

- \( \theta^T T(x) \) is a real-valued “score” (positive or negative), defined in terms of “features” \( T(x) \) and parameters \( \theta \)
- \( \exp(\theta^T T(x)) \) is an unnormalized probability
- The log-partition \( A(\theta) = \log Z(\theta) \) function ensures normalization

\[ p_\theta(x) = \frac{\exp(\theta^T T(x))}{\exp(A(\theta))}, \quad A(\theta) = \log Z(\theta) = \log \int \exp(\theta^T T(x)) \, dx \]

- Valid parameters are the ones for which \( A(\theta) \) is finite.

Applications and Importance

- We can get many different families of distributions by selecting different “features” \( T(x) \) for a variable \( x \) in some sample space:
  - Bernoulli, Binomial, Multinomial, Beta, Gaussian, Poisson, MRFs, ...
- There is a general theory that covers learning and other properties of all of these distributions!
- A good trick to seeing that a distribution belongs to an exponential family is to match its log-density to

\[ \log p_\theta(x) = \log h(x) + \theta^T T(x) - A(\theta) \]
Preview: Graphical Models

For some intuition why exponential families could be relevant for graphical models, observe that the unnormalized probability factors over "simpler" functions, just like graphical models:

$$\exp(\theta^T T(x)) = \exp \sum_i \theta_i T_i(x) = \prod_i \exp(\theta_i T_i(x))$$

(Think: what could $T(x)$ look like to recover a graphical model?)

Example: Bernoulli Distribution

The Bernoulli distribution with parameter $\mu \in [0,1]$ has density (pmf)

$$p_{\mu}(x) = \begin{cases} 
\mu & x = 1 \\
1 - \mu & x = 0 
\end{cases}$$

One way to write the log-density is

$$\log p_{\mu}(x) = [I[x = 1] \log \mu + I[x = 0] \log(1 - \mu)]$$

To match this to an exponential family

$$\log p_{\theta}(x) = \log h(x) + \theta^T T(x) - A(\theta),$$

where

$$T(x) = (I[x = 1], I[x = 0])$$

$$\Theta = (\log \mu, \log(1 - \mu)) \in \mathbb{R}^2$$

$$h(x) = 1$$

$$T(x) = (I[x = 1], I[x = 0])$$

$$\theta = (\log \mu, \log(1 - \mu))$$

$$\exp(\theta^T T(x)) = \begin{cases} 
\exp(\theta_1) & x = 1 \\
\exp(\theta_2) & x = 0 
\end{cases}$$

$$A(\theta) = \log(\exp(\theta_1) + \exp(\theta_2))$$

$$A(\theta) = (\log \mu, \log(1 - \mu))$$

It's easy to check that $A(\theta) = 0$ when $\theta = (\log \mu, \log(1 - \mu))$.

Review: Bernoulli Distribution

To match this to an exponential family $\log p_{\theta}(x) = \log h(x) + \theta^T T(x) - A(\theta)$, take

- $h(x) = 1$
- $T(x) = (I[x = 1], I[x = 0])$
- $\theta = (\log \mu, \log(1 - \mu))$
- $\exp(\theta^T T(x)) = \begin{cases} 
\exp(\theta_1) & x = 1 \\
\exp(\theta_2) & x = 0 
\end{cases}$
- $A(\theta) = \log(\exp(\theta_1) + \exp(\theta_2))$
- It's easy to check that $A(\theta) = 0$ when $\theta = (\log \mu, \log(1 - \mu))$.
Example: Bernoulli, Single Parameter

We can also write the Bernoulli as a single-parameter exponential family. Rewrite the log-density as

$$\log p_\mu(x) = \log(1 - \mu) + x \log \frac{\mu}{1 - \mu}.$$
Pairwise Markov Random Field

Will revisit later...