Review:
Probability and Statistics

Sam Roweis

October, 2002

Random Variables and Densities

• Random variables X represents outcomes or states of world. Instantiations of variables usually in lower case: x. We will write \( p(x) \) to mean probability(\( X = x \)).
• Sample Space: the space of all possible outcomes/states. (May be discrete or continuous or mixed.)
• Probability mass (density) function \( p(x) \geq 0 \) Assigns a non-negative number to each point in sample space. Sums (integrates) to unity: \( \sum_x p(x) = 1 \) or \( \int_x p(x)dx = 1. \) Intuitively: how often does \( x \) occur, how much do we believe in \( x \).
• Ensemble: random variable + sample space+ probability function

But you have to be careful about the context and about defining random variables and sample spaces carefully. Otherwise you can get in trouble (see, e.g. Simpson’s paradox/Prisoner’s paradox).

Probability

• We use probabilities \( p(x) \) to represent our beliefs \( B(x) \) about the states \( x \) of the world.
• There is a formal calculus for manipulating uncertainties represented by probabilities.
• Any consistent set of beliefs obeying the Cox Axioms can be mapped into probabilities.
  1. Rationally ordered degrees of belief:
     if \( B(x) > B(y) \) and \( B(y) > B(z) \) then \( B(x) > B(z) \)
  2. Belief in \( x \) and its negation \( \bar{x} \) are related: \( B(x) = f[B(\bar{x})] \)
  3. Belief in conjunction depends only on conditionals:
     \( B(x \text{ and } y) = g[B(x), B(y|x)] = g[B(y), B(x|y)] \)

Expectations, Moments

• Expectation of a function \( a(x) \) is written \( E[a] \) or \( \langle a \rangle \)

\[
E[a] = \langle a \rangle = \sum_x p(x)a(x)
\]

E.g. mean = \( \sum_x xp(x) \), variance = \( \sum_x (x - E[x])^2 p(x) \)
• Moments are expectations of higher order powers.
  (Mean is first moment. Autocorrelation is second moment.)
• Centralized moments have lower moments subtracted away (e.g. variance, skew, curtoysis).
• Deep fact: Knowledge of all orders of moments completely defines the entire distribution.
**Joint Probability**

- Key concept: two or more random variables may interact. Thus, the probability of one taking on a certain value depends on which value(s) the others are taking.
- We call this a joint ensemble and write
  \[ p(x, y) = \text{prob}(X = x \text{ and } Y = y) \]

**Marginal Probabilities**

- We can "sum out" part of a joint distribution to get the marginal distribution of a subset of variables:
  \[ p(x) = \sum_y p(x, y) \]
- This is like adding slices of the table together.
- Another equivalent definition: \( p(x) = \sum_y p(x | y)p(y) \).

**Conditional Probability**

- If we know that some event has occurred, it changes our belief about the probability of other events.
- This is like taking a "slice" through the joint table.
  \[ p(x | y) = \frac{p(x, y)}{p(y)} \]

**Bayes’ Rule**

- Manipulating the basic definition of conditional probability gives one of the most important formulas in probability theory:
  \[ p(x | y) = \frac{p(y | x)p(x)}{p(y)} = \frac{p(y | x)p(x)}{\sum_{x'} p(y | x')p(x')} \]
- This gives us a way of "reversing" conditional probabilities.
- Thus, all joint probabilities can be factored by selecting an ordering for the random variables and using the "chain rule":
  \[ p(x, y, z, \ldots) = p(x)p(y | x)p(z | x, y)p(\ldots | x, y, z) \]
**Independence & Conditional Independence**

- Two variables are independent iff their joint factors:
  \[ p(x, y) = p(x)p(y) \]

- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:
  \[ p(x, y | z) = p(x | z)p(y | z) \quad \forall z \]

**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 \(- \log p(event)\)

**Cross Entropy (KL Divergence)**

- An asymmetric measure of the distance between two distributions:
  \[ KL[p || q] = \sum_x p(x)[\log p(x) - \log q(x)] \]

- \(KL > 0\) 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 \(q(x)\).

**Jensen’s Inequality**

- For any concave function \(f()\) and any distribution on \(x\),
  \[ E[f(x)] \leq f(E[x]) \]

- e.g. \(\log()\) and \(\sqrt{}\) are concave
- This allows us to bound expressions like \(\log p(x) = \log \sum_z p(x, z)\)
Statistics

- Probability: inferring probabilistic quantities for data given fixed models (e.g. prob. of events, marginals, conditionals, etc).
- Statistics: inferring a model given fixed data observations (e.g. clustering, classification, regression).
- Many approaches to statistics: frequentist, Bayesian, decision theory, ...

(Conditional) Probability Tables

- For discrete (categorical) quantities, the most basic parametrization is the probability table which lists \( p(x = k^{th} \text{ value}) \).
- Since PTs must be nonnegative and sum to 1, for \( k \)-ary variables there are \( k - 1 \) free parameters.
- If a discrete variable is conditioned on the values of some other discrete variables we make one table for each possible setting of the parents: these are called conditional probability tables or CPTs.

Some (Conditional) Probability Functions

- Probability density functions \( p(x) \) (for continuous variables) or probability mass functions \( p(x = k) \) (for discrete variables) tell us how likely it is to get a particular value for a random variable (possibly conditioned on the values of some other variables.)
- We can consider various types of variables: binary/discrete (categorical), continuous, interval, and integer counts.
- For each type we'll see some basic probability models which are parametrized families of distributions.

Exponential Family

- For (continuous or discrete) random variable \( x \)
  \[
p(x|\eta) = h(x) \exp\{\eta^\top T(x) - A(\eta)\}
  = \frac{1}{Z(\eta)} h(x) \exp\{\eta^\top T(x)\}
\]
is an exponential family distribution with natural parameter \( \eta \).
- Function \( T(x) \) is a sufficient statistic.
- Function \( A(\eta) = \log Z(\eta) \) is the log normalizer.
- Key idea: all you need to know about the data is captured in the summarizing function \( T(x) \).
### Bernoulli

- For a binary random variable with \( p(\text{heads}) = \pi \):
  
  \[
  p(x | \pi) = \pi^x (1 - \pi)^{1-x} = \exp \left\{ \log \left( \frac{\pi}{1 - \pi} \right) x + \log(1 - \pi) \right\}
  \]

- Exponential family with:
  
  \[
  \eta = \log \frac{\pi}{1 - \pi} \\
  T(x) = x \\
  A(\eta) = -\log(1 - \pi) = \log(1 + e^{\eta}) \\
  h(x) = 1
  \]

- The logistic function relates the natural parameter and the chance of heads
  
  \[
  \pi = \frac{1}{1 + e^{-\eta}}
  \]

### Poisson

- For an integer count variable with rate \( \lambda \):
  
  \[
  p(x | \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} \exp \{ x \log \lambda - \lambda \}
  \]

- Exponential family with:
  
  \[
  \eta = \log \lambda \\
  T(x) = x \\
  A(\eta) = \lambda = e^\eta \\
  h(x) = \frac{1}{x!}
  \]

- e.g. number of photons \( x \) that arrive at a pixel during a fixed interval given mean intensity \( \lambda \)
- Other count densities: binomial, exponential.

### Multinomial

- For a set of integer counts on \( k \) trials
  
  \[
  p(x | \pi) = \frac{k!}{x_1! x_2! \cdots x_n!} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_n^{x_n} = h(x) \exp \left\{ \sum_i x_i \log \pi_i \right\}
  \]

- But the parameters are constrained: \( \sum_i \pi_i = 1 \).
  
  So we define the last one \( \pi_n = 1 - \sum_{i=1}^{n-1} \pi_i \).

  \[
  p(x | \pi) = h(x) \exp \left\{ \sum_{i=1}^{n-1} \log \left( \frac{\pi_i}{\pi_n} \right) x_i + k \log \pi_n \right\}
  \]

- Exponential family with:
  
  \[
  \eta_i = \log \pi_i - \log \pi_n \\
  T(x_i) = x_i \\
  A(\eta) = -k \log \pi_n = k \log \sum_i e^{\eta_i} \\
  h(x) = \frac{k!}{x_1! x_2! \cdots x_n!}
  \]

  The softmax function relates the basic and natural parameters:

  \[
  \pi_i = \frac{e^{\eta_i}}{\sum_j e^{\eta_j}}
  \]

### Gaussian (Normal)

- For a continuous univariate random variable:
  
  \[
  p(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log \sigma \right\}
  \]

- Exponential family with:
  
  \[
  \eta = [\mu/\sigma^2; -1/2\sigma^2] \\
  T(x) = [x; x^2] \\
  A(\eta) = \log \sigma + \mu/2\sigma^2 \\
  h(x) = 1/\sqrt{2\pi}
  \]

- Note: a univariate Gaussian is a two-parameter distribution with a two-component vector of sufficient statistics.
**Multivariate Gaussian Distribution**

- For a continuous vector random variable:
  \[
p(x|\mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}
  \]

- Exponential family with:
  \[
  \eta = [\Sigma^{-1} \mu ; -1/2 \Sigma^{-1}]
  \]
  \[
  T(x) = [x ; xx^\top]
  \]
  \[
  A(\eta) = \log |\Sigma|/2 + \mu^\top \Sigma^{-1} \mu/2
  \]
  \[
  h(x) = (2\pi)^{-n/2}
  \]

- Sufficient statistics: mean vector and correlation matrix.
- Other densities: Student-t, Laplacian.
- For non-negative values use exponential, Gamma, log-normal.

**Gaussian Marginals/Conditionals**

- To find these parameters is mostly linear algebra:
  Let \( z = [x^\top y^\top]^\top \) be normally distributed according to:
  \[
  z = \begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} a \\ b \end{bmatrix}; \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix} \right)
  \]
  where \( C \) is the (non-symmetric) cross-covariance matrix between \( x \) and \( y \) which has as many rows as the size of \( x \) and as many columns as the size of \( y \).
  The marginal distributions are:
  \[
  x \sim \mathcal{N}(a; A)
  \]
  \[
  y \sim \mathcal{N}(b; B)
  \]
  and the conditional distributions are:
  \[
  x|y \sim \mathcal{N}(a + CB^{-1}(y - b); A - CB^{-1}C^\top)
  \]
  \[
  y|x \sim \mathcal{N}(b + C^\top A^{-1}(x - a); B - C^\top A^{-1}C)
  \]

**Important Gaussian Facts**

- All marginals of a Gaussian are again Gaussian.
- Any conditional of a Gaussian is again Gaussian.

**Moments**

- For continuous variables, moment calculations are important.
- We can easily compute moments of any exponential family distribution by taking the derivatives of the log normalizer \( A(\eta) \).
- The \( q^{th} \) derivative gives the \( q^{th} \) centred moment.
  \[
  \frac{dA(\eta)}{d\eta} = \text{mean}
  \]
  \[
  \frac{d^2A(\eta)}{d\eta^2} = \text{variance}
  \]
  \[
  \ldots
  \]
- When the sufficient statistic is a vector, partial derivatives need to be considered.
**Linear-Gaussian Conditionals**

- When the variable(s) being conditioned on (parents) are discrete, we just have one density for each possible setting of the parents. e.g. a table of natural parameters in exponential models or a table of tables for discrete models.
- When the conditioned variable is continuous, its value sets some of the parameters for the other variables.
- A very common instance of this for regression is the “linear-Gaussian”: 
  \[ p(y|x) = \text{gauss}(\theta^\top x; \Sigma). \]
- For discrete children and continuous parents, we often use a Bernoulli/multinomial whose parameters are some function \( f(\theta^\top x) \).

**Likelihood Function**

- So far we have focused on the (log) probability function \( p(x|\theta) \) which assigns a probability (density) to any joint configuration of variables \( x \) given fixed parameters \( \theta \).
- But in learning we turn this on its head: we have some fixed data and we want to find parameters.
- Think of \( p(x|\theta) \) as a function of \( \theta \) for fixed \( x \):
  \[
  L(\theta; x) = p(x|\theta) \\
  \ell(\theta; x) = \log p(x|\theta)
  \]
  This function is called the (log) “likelihood”.
- Chose \( \theta \) to maximize some cost function \( c(\theta) \) which includes \( \ell(\theta) \):
  \[
  c(\theta) = \ell(\theta; \mathcal{D}) \quad \text{maximum likelihood (ML)} \\
  c(\theta) = \ell(\theta; \mathcal{D}) + r(\theta) \quad \text{maximum a posteriori (MAP)/penalizedML}
  \]
  (also cross-validation, Bayesian estimators, BIC, AIC, ...)

**Multiple Observations, Complete Data, IID Sampling**

- A single observation of the data \( x \) is rarely useful on its own.
- Generally we have data including many observations, which creates a set of random variables: \( \mathcal{D} = \{x^1, x^2, \ldots, x^M\} \)
- Two very common assumptions:
  1. Observations are independently and identically distributed (IID) according to joint distribution of graphical model: IID samples.
  2. We observe all random variables in the domain on each observation: complete data.

**Maximum Likelihood**

- For IID data:
  \[
  p(\mathcal{D}|\theta) = \prod_m p(x^m|\theta) \\
  \ell(\theta; \mathcal{D}) = \sum_m \log p(x^m|\theta)
  \]
- Idea of maximum likelihood estimation (MLE): pick the setting of parameters most likely to have generated the data we saw:
  \[
  \theta_{\text{ML}}^* = \text{argmax}_\theta \ell(\theta; \mathcal{D})
  \]
- Very commonly used in statistics. Often leads to “intuitive”, “appealing”, or “natural” estimators.
**Example: Bernoulli Trials**

- We observe $M$ iid coin flips: $D = H, H, T, H, \ldots$
- Model: $p(H) = \theta \quad p(T) = (1 - \theta)$
- Likelihood:
  \[
  \ell(\theta; D) = \log p(D|\theta) = \log \prod_m \theta^{x_m} (1 - \theta)^{1-x_m}
  = \log \theta \sum_m x_m + \log(1 - \theta) \sum_m (1 - x_m)
  = \log \theta N_H + \log(1 - \theta) N_T
  \]
- Take derivatives and set to zero:
  \[
  \frac{\partial \ell}{\partial \theta} = N_H \theta - N_T (1 - \theta)
  \Rightarrow \theta_{\text{ML}}^* = \frac{N_H}{N_H + N_T}
  \]

---

**Example: Multinomial**

- We observe $M$ iid die rolls (K-sided): $D = 3, 1, K, 2, \ldots$
- Model: $p(k) = \theta_k \quad \sum_k \theta_k = 1$
- Likelihood (for binary indicators $[x^m = k]$):
  \[
  \ell(\theta; D) = \log p(D|\theta) = \log \prod_m \theta_{x^m} = \log \prod_m \theta_{[x^m=1]} \ldots \theta_{[x^m=k]}
  = \sum_k \log \theta_k \sum_m [x^m = k] = \sum_k N_k \log \theta_k
  \]
- Take derivatives and set to zero (enforcing $\sum_k \theta_k = 1$):
  \[
  \frac{\partial \ell}{\partial \theta_k} = \frac{N_k}{\theta_k} - M
  \Rightarrow \theta_{k,\text{ML}}^* = \frac{N_k}{M}
  \]

---

**Example: Univariate Normal**

- We observe $M$ iid real samples: $D = 1.18, -0.25, 0.78, \ldots$
- Model: $p(x) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$
- Likelihood (using probability density):
  \[
  \ell(\theta; D) = \log p(D|\theta) = -\frac{M}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_m \frac{(x^m - \mu)^2}{\sigma^2}
  \]
- Take derivatives and set to zero:
  \[
  \frac{\partial \ell}{\partial \mu} = (1/\sigma^2) \sum_m (x^m - \mu)
  \frac{\partial \ell}{\partial \sigma^2} = -M/2\sigma^2 + \frac{1}{2\sigma^4} \sum_m (x^m - \mu)^2
  \Rightarrow \mu_{\text{ML}} = (1/M) \sum_m x^m
  \sigma_{\text{ML}}^2 = (1/M) \sum_m x^m - \mu_{\text{ML}}^2
  \]
**Example: Linear Regression**

- In linear regression, some inputs (covariates, parents) and all outputs (responses, children) are continuous valued variables.
- For each child and setting of discrete parents we use the model:
  \[ p(y|x, \theta) = \text{gauss}(y|\theta^\top x, \sigma^2) \]
- The likelihood is the familiar “squared error” cost:
  \[
  \ell(\theta; D) = -\frac{1}{2\sigma^2} \sum_m (y^m - \theta^\top x^m)^2
  \]
- The ML parameters can be solved for using linear least-squares:
  \[
  \frac{\partial \ell}{\partial \theta} = -\sum_m (y^m - \theta^\top x^m)x^m
  \Rightarrow \theta^*_{\text{ML}} = (X^\top X)^{-1}X^\top Y
  \]

**Sufficient Statistics**

- A statistic is a function of a random variable.
- \( T(X) \) is a “sufficient statistic” for \( X \) if
  \[
  T(x^1) = T(x^2) \Rightarrow L(\theta; x^1) = L(\theta; x^2) \quad \forall \theta
  \]
- Equivalently (by the Neyman factorization theorem) we can write:
  \[
  p(x|\theta) = h(x, T(x))g(T(x), \theta)
  \]
- Example: exponential family models:
  \[
  p(x|\theta) = h(x) \exp\{\eta^\top T(x) - A(\eta)\}
  \]

**Example: Linear Regression**

**Sufficient Statistics are Sums**

- In the examples above, the sufficient statistics were merely sums (counts) of the data:
  - Bernoulli: # of heads, tails
  - Multinomial: # of each type
  - Gaussian: mean, mean-square
  - Regression: correlations
- As we will see, this is true for all exponential family models: sufficient statistics are average natural parameters.
- Only exponential family models have simple sufficient statistics. (There are some degenerate exceptions, e.g. the uniform has sufficient statistics of max/min.)
MLE for Exponential Family Models

- Recall the probability function for exponential models:
  \[ p(x|\theta) = h(x) \exp\{\eta^\top T(x) - A(\eta)\} \]
- For iid data, sufficient statistic is \( \sum_m T(x^m) \):
  \[ \ell(\eta; D) = \log p(D|\eta) = \left( \sum_m \log h(x^m) \right) - MA(\eta) + \left( \eta^\top \sum_m T(x^m) \right) \]
- Take derivatives and set to zero:
  \[ \frac{\partial \ell}{\partial \eta} = \sum_m T(x^m) - M \frac{\partial A(\eta)}{\partial \eta} \]

  \[ \Rightarrow \frac{\partial A(\eta)}{\partial \eta} = \frac{1}{M} \sum_m T(x^m) \]

  \[ \eta_{ML} = \frac{1}{M} \sum_m T(x^m) \]

recalling that the natural moments of an exponential distribution are the derivatives of the log normalizer.

Fundamental Operations with Distributions

- Generate data: draw samples from the distribution. This often involves generating a uniformly distributed variable in the range [0,1] and transforming it. For more complex distributions it may involve an iterative procedure that takes a long time to produce a single sample (e.g. Gibbs sampling, MCMC).
- Compute log probabilities.
  When all variables are either observed or marginalized the result is a single number which is the log prob of the configuration.
- Inference: Compute expectations of some variables given others which are observed or marginalized.
- Learning:
  Set the parameters of the density functions given some (partially) observed data to maximize likelihood or penalized likelihood.

Basic Statistical Problems

- Let’s remind ourselves of the basic problems we discussed on the first day: density estimation, clustering classification and regression.
- Density estimation is hardest. If we can do joint density estimation then we can always condition to get what we want:
  Regression: \[ p(y|x) = p(y, x)/p(x) \]
  Classification: \[ p(c|x) = p(c, x)/p(x) \]
  Clustering: \[ p(c|x) = p(c, x)/p(x) \] c unobserved

Learning with Known Model Structure

- In AI the bottleneck is often knowledge acquisition.
- Human experts are rare, expensive, unreliable, slow.
- But we have lots of data.
- Want to build systems automatically based on data and a small amount of prior information (from experts).
- Many systems we build will be essentially probability models.
- Assume the prior information we have specifies type & structure of the model, as well as the form of the (conditional) distributions or potentials.
- In this case learning \( \equiv \) setting parameters.
- Also possible to do “structure learning” to learn model.