Review:

Probability and Statistics

Sam Roweis

Machine Learning Summer School, January 2005

---

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

---

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:
     \[ \text{if } B(x) > B(y) \text{ and } B(y) > B(z) \Rightarrow B(x) > B(z) \]
  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, curtosis).
- Deep fact: Knowledge of all orders of moments completely defines the entire distribution.
**Means, Variances and Covariances**

- Remember the definition of the mean and covariance of a vector random variable:
  \[ E[x] = \int x p(x) dx = m \]
  \[ \text{Cov}[x] = E[(x - m)(x - m)^\top] = \int (x - m)(x - m)^\top p(x) dx = V \]
  which is the expected value of the outer product of the variable with itself, after subtracting the mean.
- Also, the covariance between two variables:
  \[ \text{Cov}[x, y] = E[(x - m_x)(y - m_y)^\top] = C \]
  \[ = \int_{xy} (x - m_x)(y - m_y)^\top p(x, y) dx dy = C \]
  which is the expected value of the outer product of one variable with another, after subtracting their means.
  Note: \( C \) is not symmetric.

**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.

**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”:

\[ p(x, y, z, \ldots) = p(x)p(y | x)p(z | x, y)p(\ldots | x, y, z) \]

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 \(-\log p(\text{event})\).

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:

\[ p(x, y | z) = p(x | z)p(y | z) \quad \forall z \]

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 \(q(x)\).
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, ...

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.

(Conditional) Probability Tables

- For discrete (categorical) quantities, the most basic parametrization is the probability table which lists $p(x_i = k^{th}$ 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.

Exponential Family

- For (continuous or discrete) random variable $x$
  \[
  p(x | \eta) = h(x) \exp \{ \eta^T T(x) - A(\eta) \} = \frac{1}{Z(\eta)} h(x) \exp \{ \eta^T T(x) \}
  \]
  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 Distribution

- For a binary random variable \( x = \{0, 1\} \) with \( p(x = 1) = \pi \):
  \[
p(x|\pi) = \pi^x (1-\pi)^{1-x}
  = \exp \left\{ \log \left( \frac{\pi}{1-\pi} \right) x + \log(1-\pi) \right\}
  \]
- Exponential family with:
  \[
  \eta = \log \frac{\pi}{1-\pi}
  T(x) = x
  A(\eta) = -\log(1-\pi) = \log(1 + \exp(\eta))
  h(x) = 1
  \]
  The logistic function links natural parameter and chance of heads
  \[
  \pi = \frac{1}{1 + \exp(-\eta)} = \text{logistic}(\eta)
  \]

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:
  \[
  \eta = \log \lambda
  T(x) = x
  A(\eta) = \lambda = \exp(\eta)
  h(x) = \frac{1}{x!}
  \]
  e.g. number of photons \( x \) that arrive at a pixel during a fixed interval given mean intensity \( \lambda \)
  Other count densities: (neg)binomial, geometric.

Multinomial

- For a categorical (discrete), random variable taking on \( K \) possible values, let \( \pi_k \) be the probability of the \( k^{th} \) value. We can use a binary vector \( x = (x_1, x_2, \ldots, x_K) \) in which \( x_k = 1 \) if and only if the variable takes on its \( k^{th} \) value. Now we can write,
  \[
p(x|\pi) = \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_K^{x_K} = \exp \left\{ \sum_i x_i \log \pi_i \right\}
  \]
  Exactly like a probability table, but written using binary vectors.
- If we observe this variable several times \( X = \{x_1, x_2, \ldots, x_N\} \), the (iid) probability depends on the total observed counts of each value:
  \[
p(X|\pi) = \prod_n p(x^n|\pi) = \exp \left\{ \sum_i (\sum_n x_{ni}^n) \log \pi_i \right\} = \exp \left\{ \sum_i \epsilon_i \log \pi_i \right\}
  \]

Multinomial as Exponential Family

- The multinomial parameters are constrained: \( \sum_i \pi_i = 1 \).
  Define (the last) one in terms of the rest: \( \pi_K = 1 - \sum_{i=1}^{K-1} \pi_i \)
  \[
p(x|\pi) = \exp \left\{ \sum_{i=1}^{K-1} \log \left( \frac{\pi_i}{\pi_K} \right) x_i + k \log \pi_K \right\}
  \]
- Exponential family with:
  \[
  \eta_i = \log \pi_i - \log \pi_K
  T(x_i) = x_i
  A(\eta) = -k \log \pi_K = k \log \sum_i \exp(\eta_i)
  h(x) = 1
  \]
  The softmax function relates direct and natural parameters:
  \[
  \pi_i = \frac{\exp(\eta_i)}{\sum_j \exp(\eta_j)}
  \]
**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.

**Important Gaussian Facts**

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

**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:
  \[
  x|y \sim \mathcal{N}(a + CB^{-1}(y - b); A - CB^{-1}C^\top) \\
  y|x \sim \mathcal{N}(b + C^\top A^{-1}(x - a); B - C^\top A^{-1}C)
  \]
**Parameter Constraints**

- If we want to use general optimizations (e.g. conjugate gradient) to learn latent variable models, we often have to make sure parameters respect certain constraints. (e.g. $\sum_k \alpha_k = 1$, $\Sigma_k$ pos.definite).
- A good trick is to reparameterize these quantities in terms of unconstrained values. For mixing proportions, use the softmax:
  $$\alpha_k = \frac{\exp(q_k)}{\sum_j \exp(q_j)}$$
- For covariance matrices, use the Cholesky decomposition:
  $$\Sigma^{-1} = A^\top A$$
  $$|\Sigma|^{-1/2} = \prod_i A_{ii}$$

  where $A$ is upper diagonal with positive diagonal:
  $$A_{ii} = \exp(r_i) > 0 \quad A_{ij} = a_{ij} \quad (j > i) \quad A_{ij} = 0 \quad (j < i)$$

**Parametersing 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 $f(\theta^\top x)$.

**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.

**Generalized Linear Models (GLMs)**

- Generalized Linear Models: $p(y|x)$ is exponential family with conditional mean $\mu = f(\theta^\top x)$.
- The function $f$ is called the response function; if we chose it to be the inverse of the mapping b/w conditional mean and natural parameters then it is called the canonical response function.
  $$\eta = \psi(\mu)$$
  $$f(\cdot) = \psi^{-1}(\cdot)$$
- We can be even more general and define distributions by arbitrary energy functions proportional to the log probability:
  $$p(x) \propto \exp\{-\sum_k H_k(x)\}$$
- A common choice is to use pairwise terms in the energy:
  $$H(x) = \sum_i a_i x_i + \sum\text{pairs } i j w_{ij} x_i x_j$$
**Matrix Inversion Lemma**  
(Sherman-Morrison-Woodbury Formulae)

- There is a good trick for inverting matrices when they can be decomposed into the sum of an easily inverted matrix \( D \) and a low rank outer product. It is called the matrix inversion lemma.
\[
(D - AB^{-1}A^\top)^{-1} = D^{-1} + D^{-1}A(B - A^\top D^{-1}A)^{-1}A^\top D^{-1}
\]
- The same trick can be used to compute determinants:
\[
\log |D + AB^{-1}A^\top| = \log |D| - \log |B| + \log |B + A^\top D^{-1}A|
\]

**Matrix Derivatives**

- Here are some useful matrix derivatives:
\[
\begin{align*}
\frac{\partial}{\partial A} \log |A| &= (A^{-1})^\top \\
\frac{\partial}{\partial A} \text{trace}[B^\top A] &= B \\
\frac{\partial}{\partial A} \text{trace}[BA^\top CA] &= 2CAB
\end{align*}
\]

**Jensen’s Inequality**

- For any concave function \( f() \) and any distribution on \( x \),
\[
E[f(x)] \leq f(E[x])
\]
- e.g. \( \log() \) and \( \sqrt{\cdot} \) are concave
- This allows us to bound expressions like \( \log p(x) = \log \sum_z p(x, z) \)

**Logsum**

- Often you can easily compute \( b_k = \log p(x|z = k, \theta_k) \), but it will be very negative, say \(-10^6\) or smaller.
- Now, to compute \( \ell = \log p(x|\theta) \) you need to compute \( \log \sum_k e^{b_k} \).
  (e.g. for calculating responsibilities at test time or for learning)
- Careful! Do not compute this by doing \( \log(\text{sum}(\exp(b))) \).
  You will get underflow and an incorrect answer.
- Instead do this:
  - Add a constant exponent \( B \) to all the values \( b_k \) such that the largest value comes close to the maximum exponent allowed by machine precision: \( B = \text{MAXEXPO\text{NENT}} - \log(K) - \max(b) \).
  - Compute \( \log(\text{sum}(\exp(b+B))) - B \).
- Example: if \( \log p(x|z = 1) = -120 \) and \( \log p(x|z = 2) = -120 \), what is \( \log p(x) = \log [p(x|z = 1) + p(x|z = 2)] \)?
  Answer: \( \log[2e^{-120}] = -120 + \log 2 \).