CSC321 Tutorial 4: Probabilities for machine learning (Most slides made by Roland Memisevic and Sam Roweis)

Yue Li
Email: yueli@cs.toronto.edu

Wed 11-12 Feb 5
Fri 10-11 Feb 7

Why probabilities?

- One of the hardest problems when building complex intelligent systems is **brittleness**.
- How can we keep tiny irregularities from causing everything to break?

Keeping all options open

- **Probabilities** are a great formalism for avoiding brittleness, because they allow us to be *explicit about uncertainties:*
- Instead of representing *values*: Define *distributions over alternatives*!
- Example: Instead of *setting* values strictly ('$x = 4$'), define all of: $p(x = 1), p(x = 2), p(x = 3), p(x = 4), p(x = 5)$
- Great success story. Most powerful machine learning models consider probabilities in some way.
- (Note that we could still *express* things like '$x = 4$'. (How?))

"Not random, not a variable."

- For p we need: $\sum_x p(x) = 1$ and $p(x) \geq 0$
- Formally, the 'object taking on random values' is called **random variable** and $p(\cdot)$ is its **distribution.**
- Capital letters ('X') often used for random variables, small letters ('x') for values it takes on.
- Sometimes we see $p(X = x)$, but usually just $p(x)$.
- In general, the symbol p is often heavily overloaded and the argument decides.
- These are notational quirks that require a little time to get used to, but make life easier later on.

Continuous random variables

- For continuous x we can replace $\sum$ by $\int$, but ...
- Things work somewhat differently for continuous x. For example, we have $p(X = \mathrm{value}) = 0$ for any value.
- Only things like $p(X \in [-0.5, 0.7])$ are reasonable.
- The reason is the integral...
- (Note, again, that p is overloaded.)

Summarizing properties

- The interesting **properties** of RVs are usually just properties of their distributions (not surprisingly).
- Mean:

$$\mu =$$

Summarizing properties

- The interesting **properties** of RVs are usually just properties of their distributions (not surprisingly).
- Mean:

$$\mu = \sum_x p(x)x$$

Summarizing properties

- The interesting **properties** of RVs are usually just properties of their distributions (not surprisingly).
- Mean:

$$\mu = \sum_x p(x)x$$

- Variance:

$$\sigma^2 =$$

Summarizing properties

- The interesting **properties** of RVs are usually just properties of their distributions (not surprisingly).
- Mean:

$$\mu = \sum_x p(x)x$$

- Variance:

$$\sigma^2 = \sum_x p(x)(x-\mu)^2$$

Summarizing properties

- The interesting **properties** of RVs are usually just properties of their distributions (not surprisingly).
- Mean:

$$\mu = \sum_x p(x)x$$

- Variance:

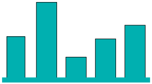
$$\sigma^2 = \sum_x p(x)(x-\mu)^2$$

- (Standard deviation: $\sigma = \sqrt{\sigma^2}$)

`Maximum likelihood estimate (MLE) of Gaussian (later slides)`

Some standard distributions

Discrete

- Multinomial..... 
- Bernoulli... $p^x(1-p)^{1-x}$ (x is zero or one)
- Binomial..... 'Sum of Bernoullis' (unfortunate naming confusion). Actually, also the multinomial is often defined as a distribution over the *sum* of outcomes of our 'multinomial' defined above.
- Poisson, uniform, geometric, ...

Continuous

- Uniform..... 
- Gaussian... $p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$
- Etc...

`distribution in more detailed ...`

Exponential Family

- For (continuous or discrete) random variable $\mathbf{x}$

$$\begin{aligned} p(\mathbf{x}|\eta) &= h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x}) - A(\eta)\} \\ &= \frac{1}{Z(\eta)} h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x})\} \end{aligned}$$

  is an exponential family distribution with *natural parameter* η.
- Function $T(\mathbf{x})$ 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 $T(\mathbf{x})$.

- For a binary random variable with p(heads)$=\pi$:

$$\begin{aligned} p(x|\pi) &= \pi^x(1-\pi)^{1-x} \\ &= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\} \end{aligned}$$

- Exponential family with:

$$\begin{aligned} \eta &= \log\frac{\pi}{1-\pi} \\ T(x) &= x \\ A(\eta) &= -\log(1-\pi) = \log(1+e^{\eta}) \\ h(x) &= 1 \end{aligned}$$

- 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 λ:

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!}\exp\{x \log\lambda - \lambda\}$$

- Exponential family with:

$$\eta = \log\lambda$$
$$T(x) = x$$
$$A(\eta) = \lambda = e^{\eta}$$
$$h(x) = \frac{1}{x!}$$

- e.g. number of photons $\mathbf{x}$ that arrive at a pixel during a fixed interval given mean intensity λ
- Other count densities: binomial, exponential.

Multinomial

- For a set of integer counts on k trials

$$p(\mathbf{x}|\pi) = \frac{k!}{x_1!x_2!\cdots x_n!}\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_n^{x_n} = h(\mathbf{x})\exp\left\{\sum_i x_i \log \pi_i\right\}$$

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

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

- Exponential family with:

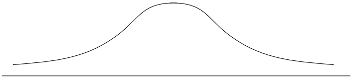
$$\begin{aligned}
\eta_i &= \log\pi_i - \log\pi_n \\
T(x_i) &= x_i \\
A(\eta) &= -k\log\pi_n = k\log\sum_i e^{\eta_i} \\
h(\mathbf{x}) &= k!/x_1!x_2!\cdots x_n!
\end{aligned}$$

Gaussian (normal)

- For a continuous univariate random variable:

$$
\begin{aligned}
p(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{\mu x}{\sigma^2}-\frac{x^2}{2\sigma^2}-\frac{\mu^2}{2\sigma^2}-\log\sigma\right\}
\end{aligned}
$$

- Exponential family with:



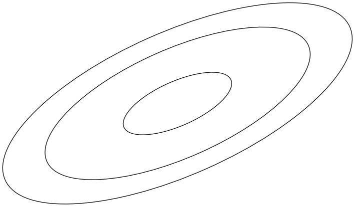
$$
\begin{aligned}
\eta &= [\mu/\sigma^2\;;\;-1/2\sigma^2] \\
T(x) &= [x\;;\;x^2] \\
A(\eta) &= \log\sigma + \mu/2\sigma^2 \\
h(x) &= 1/\sqrt{2\pi}
\end{aligned}
$$

- Note: a univariate Gaussian is a two-parameter distribution with a two-component vector of sufficient statistis.

Multivariate Gaussian Distribution

- For a continuous vector random variable:

$$p(x|\mu,\Sigma) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$



- Exponential family with:

$$\begin{aligned} \eta &= [\Sigma^{-1}\mu\ ;\ -1/2\Sigma^{-1}] \\ T(x) &= [\mathbf{x}\ ;\ \mathbf{x}\mathbf{x}^\top] \\ A(\eta) &= \log|\Sigma|/2 + \mu^\top\Sigma^{-1}\mu/2 \\ h(x) &= (2\pi)^{-n/2} \end{aligned}$$

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

Joints, conditionals, marginals

- Things get much more interesting if we allow for **multiple variables.**
- Leads to several new concepts:
- The **joint distribution** $p(x, y)$ is just a distribution defined on vectors (here 2-d as example)...
- For discrete RVs, we can imagine a *table*.
- Everything else stays essentially the same. So in particular we need

$$\sum_{x,y} p(x, y) = 1, \quad p(x, y) \geq 0$$

`conditional table example on board`

Joints, conditionals, marginals

- All we need to know about a random vector can be derived from the joint distribution. For example:
- **Marginal distributions**:

$$p(x) = \sum_y p(x,y) \quad \text{and} \quad p(y) = \sum_x p(x,y)$$

- Intuition: Collapse dimensions.
- **Conditional distributions** are defined as:

$$p(y|x) = \frac{p(x,y)}{p(x)} \quad \text{and} \quad p(x|y) = \frac{p(x,y)}{p(y)}$$

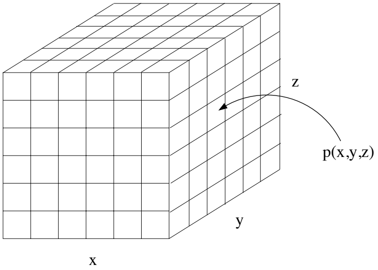
- Intuition: New frame of reference.

example from the
same table on board

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

$$p(x, y) = \text{prob}(X = x \text{ and } Y = y)$$

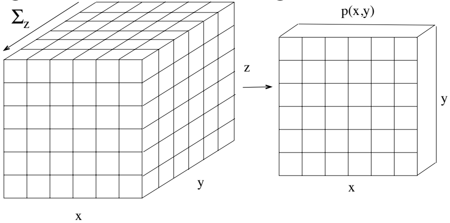


Marginal Probabilities

- 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.

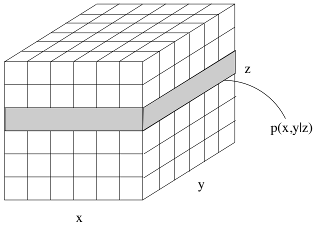


- Another equivalent definition: $p(x) = \sum_y p(x|y)p(y)$.

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.

$$\boxed{p(x|y) = p(x,y)/p(y)}$$



Important formula

- Remember this:

$$p(y|x)p(x) = p(x, y) = p(x|y)p(y)$$

- Allows us, among other things, to compute $p(x|y)$ from $p(y|x)$ ('Bayes rule').
- Can be generalized to more variables. ('Chain-rule of probability').

```
also p(x) can be defined as some known
distribution such as Gaussian
```

Bayes' Rule

- Manipulating the basic definition of conditional probability gives one of the most important formulas in probability theory:

$$\boxed{p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\sum_{x'} p(y|x')p(x')}}$$

- 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":

$$\boxed{p(x, y, z, \ldots) = p(x)p(y|x)p(z|x, y)p(\ldots|x, y, z)}$$

Independence and conditional independence

- Two RVs are called **independent**, if

$$p(x, y) = p(x)p(y)$$

what is p(x,y) when x and y are not independent? (prev. slide)

- Captures the intuition of 'independence':
- Note, for example, that it implies $p(x) = p(x|y)$.
- Related concept: x, y are called **conditionally** independent, given z if

$$p(x, y|z) = p(x|z)p(y|z)$$

see figure in next slide

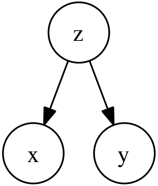
Independence is useful

- Say, we have some variables $x_1, x_2, \ldots, x_K$.
- Even just *defining* their joint (let alone doing computations with it) is hopeless for large K. O(C^K)
- But what if all x_i independent? the other extreme
- Need to specify just K probabilities, since the joint is the product! O(K)
- A more sophisticated version of this idea is to use *conditional* independence. Large and active area of 'Graphical Models'.

see figure and example

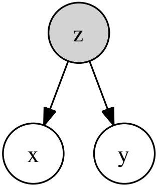
Graphical model

Conditional independence



If z has *not* been observed, x and y are in general *not* independent:

$$x \not\perp\!\!\!\perp y$$
$$p(x|y) \neq p(x)$$
$$p(x, y) \neq p(x)p(y)$$



Once z has been observed, x and y become conditionally independent:

$$x \perp\!\!\!\perp y|z$$
$$p(x|y, z) = p(x|z)$$
$$p(x, y|z) = p(x|z)p(y|z)$$

- Example 1: Suppose X and Y are the outcomes (Heads or Tails) of two separate tosses of the same coins. Clearly, X and Y are independent: $X \perp\!\!\!\perp Y$.
- Example 2: Now suppose there is a probability Z that the coin is biased towards Heads. In this case, X and Y are *not* independent: $X \not\perp\!\!\!\perp Y$.
  - Because observing that Y is Heads causes us to increase our belief in X being Heads:
    $p(X = Heads|Y = Heads) > p(X = Heads)$.
  - However, once we know such probability Z, then any evidence about Y cannot change our belief about X:
    $p(X|Z) = p(X|Y, Z)$
  - Thus, X and Y are conditionally independent given Z.

Maximum Likelihood

- Another useful thing about independence.
- Task: Given some data $(x_1, \ldots, x_N)$ build a *model* of the data-generating process. Useful for classification, novelty detection, 'image manipulation', and countless other things.
- Possible solution: Fit a **parameterized model** $p(x; w)$ to the data.
- How? Maximize the probability of 'seeing' the data under your model!

Maximum Likelihood

- This is easy, if the examples are independent, ie. if

$$p(x_1, \ldots, x_N; w) = \prod_i p(x_i; w)$$

- Note that instead of maximizing probability, we might as well maximize log probability. (Since the 'log' is monotonous.)
- So we can maximize:

$$L(w) = \log \prod_i p(x_i; w) = \sum_i \log p(x_i; w)$$

- Dealing with the sum of things is easy. (We wouldn't have gotten this, if we hadn't assumed independence.)

Example: Bernoulli Trials

- We observe M iid coin flips: $\mathcal{D}$=H,H,T,H,...
- Model: $p(H) = \theta \quad p(T) = (1 - \theta)$
- Likelihood:

$$\begin{aligned}
\ell(\theta; \mathcal{D}) &= \log p(\mathcal{D}|\theta) \\
&= \log \prod_m \theta^{\mathbf{x}^m} (1-\theta)^{1-\mathbf{x}^m} \\
&= \log \theta \sum_m \mathbf{x}^m + \log(1-\theta) \sum_m (1 - \mathbf{x}^m) \\
&= \log \theta N_{\mathrm{H}} + \log(1-\theta) N_{\mathrm{T}}
\end{aligned}$$

- Take derivatives and set to zero:

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta} &= \frac{N_{\mathrm{H}}}{\theta} - \frac{N_{\mathrm{T}}}{1-\theta} \\
\Rightarrow \theta^*_{\mathrm{ML}} &= \frac{N_{\mathrm{H}}}{N_{\mathrm{H}} + N_{\mathrm{T}}}
\end{aligned}$$

Example: Multinomial

- We observe M iid die rolls (K-sided): $\mathcal{D}$=3,1,K,2,...
- Model: $p(k) = \theta_k \quad \sum_k \theta_k = 1$
- Likelihood (for binary indicators $[\mathbf{x}^m = k]$):

$$\begin{aligned}\ell(\theta;\mathcal{D}) &= \log p(\mathcal{D}|\theta) \\ &= \log \prod_m \theta_{\mathbf{x}^m} = \log \prod_m \theta_1^{[\mathbf{x}^m=1]} \ldots \theta_k^{[\mathbf{x}^m=k]} \\ &= \sum_k \log \theta_k \sum_m [\mathbf{x}^m = k] = \sum_k N_k \log \theta_k\end{aligned}$$

- Take derivatives and set to zero (enforcing $\sum_k \theta_k = 1$):

$$\begin{aligned}\frac{\partial \ell}{\partial \theta_k} &= \frac{N_k}{\theta_k} - M \\ \Rightarrow \theta_k^* &= \frac{N_k}{M}\end{aligned}$$

Gaussian example

- What is the ML-estimate of the **mean** of a Gaussian?
- We need to maximize:

$$L(\mu) = \sum_i \log p(x_i; \mu) = \sum_i \left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) + \text{const.}$$

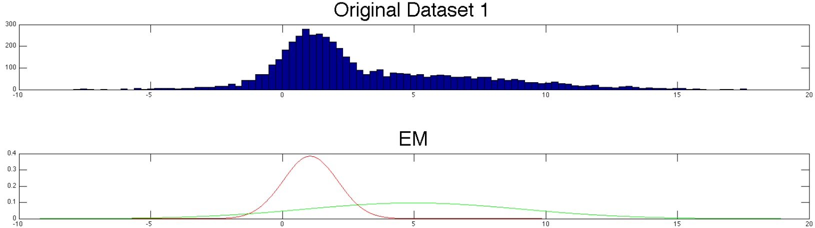
- The derivative is:

$$\frac{\partial L(\mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum_i (x_i - \mu) = \frac{1}{\sigma^2} (\sum_i x_i - N\mu)$$

- We set to zero and get:

$$\mu = \frac{1}{N} \sum_i x_i$$

When data from K Gaussians - Gaussian Mixture Model



$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \tag{1}$$

$$\ln p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \right\}. \tag{2}$$

In the E-step, the posterior probability (or $\gamma(z_k)$ as the *responsibility* of z_k for $\mathbf{x}$) is estimated as:

$$\gamma(z_k) = p(z_k|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^{old}, \Sigma_k^{old})}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^{old}, \Sigma_k^{old})} \tag{3}$$

In the M-step, the parameters involved in (1) are re-estimated by

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_k)\mathbf{x_n} \tag{4}$$

$$\Sigma_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_k)(\mathbf{x}-\boldsymbol{\mu}_k^{new})(\mathbf{x}-\boldsymbol{\mu}_k^{new})^T \tag{5}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{6}$$

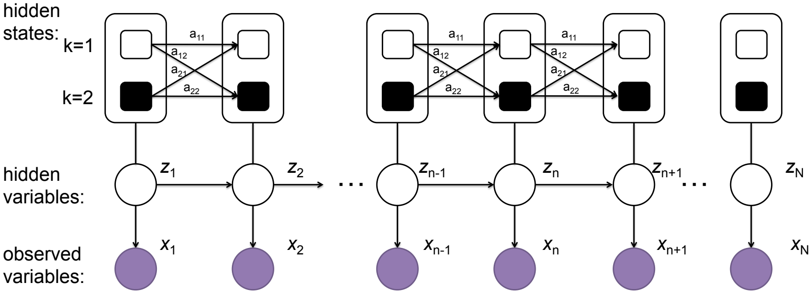
where

$$N_k = \sum_{n=1}^{N}\gamma(z_k) \tag{7}$$

The log-likelihood is then updated by

$$\ln p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k^{new}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^{new},\Sigma_k^{new})\right\}. \tag{8}$$

When data is not *i.i.d.* - Hidden Markov Model



$p(\mathbf{X}|\mathbf{Z},\phi) = \prod_{n=1}^{N} p(x_n|z_n,\phi)$ (e.g., $p(x_n|z_{n,k},\phi) = \mathcal{N}(x_i|\mu_k,\sigma_k^2)$)

$$
\begin{aligned}
p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) &= p(x_1,x_2,\ldots,x_N,z_1,z_2,\ldots,z_N|\boldsymbol{\theta}) \\
&= p(z_1|\boldsymbol{\pi})\left[\prod_{n=2}^{N} p(z_n|z_{n-1},\mathbf{A})\right]\prod_{m=1}^{N} p(x_m|z_m,\phi)
\end{aligned}
$$

Dynamic programming to obtain hidden sequence such that

$\max_{\mathbf{Z}}\left[\ln p(\mathbf{X},\mathbf{Z}|\theta)\right]$