

## Probabilities for machine learning

Roland Memisevic

Oct 19, 2011



## Complex systems and brittleness

- ▶ **Brittleness** is one of the hardest problems when building complex intelligent systems.
- ▶ 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*, we “keep all options open”: Define *distributions over alternatives*!
- ▶ Example: Instead of *setting* 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?))



## Random variables

- ▶ For  $p$  we need:  $\sum_x p(x) = 1$  and  $p(x) > 0$
- ▶ We call  $x$  **random variable** and  $p(x)$  its **distribution**.
- ▶ “Not random, not a variable.”
- ▶ The symbol  $p$  is often heavily overloaded. The argument decides.
- ▶ Also:  $p(x)$  actually short for  $p(x = \dots)$
- ▶ Notational quirks that require a little bit of time to get used to, but make life much easier later on.
- ▶ Sometimes writing  $X$  for the RV and  $x$  for values it can take on helps.
- ▶ For continuous  $x$  we can replace  $\sum$  by  $\int$ , but ...



## Continuous random variables

- ▶ Things work somewhat differently for continuous  $x$ . For example  $p(x = \text{value}) = 0$  for any value.
- ▶ Only things like  $p(x \in [-0.5, 0.7])$  are reasonable.
- ▶ Reason: we have to integrate.
- ▶ (Again: Note that  $p$  is overloaded...)



## Summarizing properties

- ▶ The interesting **properties** of RVs are just properties of their distributions (not surprisingly, since "not a variable" ...)

- ▶ Mean:

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

- ▶ Variance:

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

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



## Some useful distributions (1d)

### Discrete

- ▶ **Bernoulli:**  $p^x(1-p)^{1-x}$   
where  $x$  is either 0 or 1.

- ▶ **Discrete distribution:**  (sometimes referred to as "multinoulli")

- ▶ **Binomial, Multinomial:** Sum over Bernoulli/Discrete. (Sometimes "Multinomial" is also used to refer to a discrete distribution!)

- ▶ **Poisson:**  $p(k) = \frac{\lambda^k \exp(-\lambda)}{k!}$

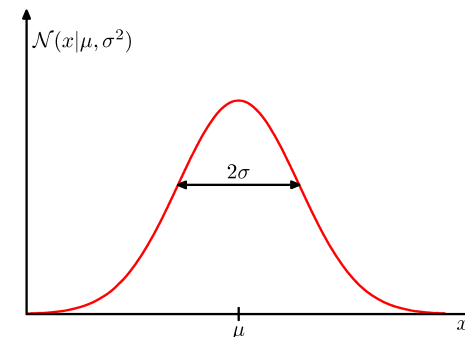
### Continuous

- ▶ **Uniform:** 

- ▶ **Gaussian (1d):**  $p(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$



## The Gaussian (1d)



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



## Joints, conditionals, marginals

- ▶ Things get really interesting only with **multiple variables**.
- ▶ Leads to several new concepts:
- ▶ The **joint distribution**  $p(x, y)$  is a function defined on more than one variable.
- ▶ For discrete RVs, imagine a *table* (or higher-dimensional *array*).
- ▶ Everything else stays the same. So, in particular we need

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



## Joints, conditionals, marginals

- ▶ All we can hope to know about a random vector can be derived from the joint distribution.
- ▶ **Marginal distributions** are given by:

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

- ▶ Obvious, when imagining tables.
- ▶ **Conditional distributions** are defined as:

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

- ▶ Think *new frame of reference*.



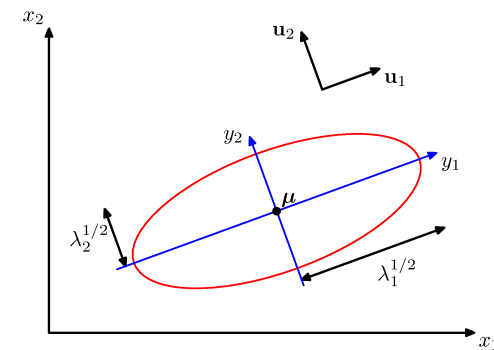
## The multivariate Gaussian

- ▶ A lot of the 1d-distributions can be generalized to higher dimensions in which case they are typically referred to as “multivariate” distributions.
- ▶ The most important multivariate distribution is the
- ▶ **Multivariate Gaussian:**

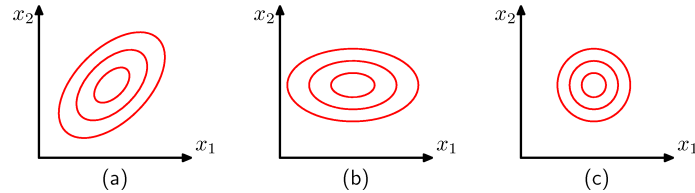
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$



## The multivariate Gaussian: Contours



## Three multivariate Gaussians



◀ ▶ ⏪ ⏩ 🔍 ↺

## A fundamental formula

- ▶ Probably the most important formula of all

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

- ▶ Can be generalized to more variables ('chainrule of probability').
- ▶ Allows us to derive **Bayes' rule**:

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

◀ ▶ ⏪ ⏩ 🔍 ↺

## Independence and conditional independence

- ▶ The two RVs are called **independent**, if

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

- ▶ Captures our intuition of "dependence".
- ▶ In particular, note that  $p(y|x) = p(y)$  follows from independence.
- ▶ Related: Two RVs are called **conditionally independent**, given a third variable  $z$ , if

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

◀ ▶ ⏪ ⏩ 🔍 ↺

## Independence is useful

- ▶ Say, we have some variables,  $x_1, x_2, \dots, x_K$
- ▶ Even just defining their joint (let alone doing computations with it) is hopeless for large  $K$  !
- ▶ But what if all the  $x_i$  are independent?
- ▶ Then we need to specify just  $K$  probabilities, because *the joint is the product!*
- ▶ A more sophisticated version of this idea is to use *conditional independence*. This builds the basis for the large and active field of *graphical models*.

◀ ▶ ⏪ ⏩ 🔍 ↺

## Maximum likelihood

- ▶ Another useful thing about independence
- ▶ Task: Given some data  $(x_1, \dots, x_N)$ , build a *model* of the data-generating process. This is the basis for many classification, novelty detection and other machine learning methods.
- ▶ Approach: Fit a **parametric model**  $p(x; \mathbf{w})$  to the data
- ▶ How? Maximize the probability of “seeing” the data under your model!



## Maximum likelihood

- ▶ This is easy if the examples are independent, that is, if:

$$p(x_1, \dots, x_N; \mathbf{w}) = \prod_i p(x_i; \mathbf{w})$$

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

$$L(\mathbf{w}) := \log \prod_i p(x_i; \mathbf{w}) = \sum_i \log p(x_i; \mathbf{w})$$



## 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:

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

- ▶ Setting to zero yields:  $\mu = \frac{1}{N} \sum_i x_i$



## Entropy

- ▶ “Probabilities allow us to be explicit about uncertainty.”
- ▶ How can we *measure* the uncertainty?
- ▶ Answer: The **entropy** of a RV (or likewise it’s distribution; remember “not a variable”) is defined as:

$$H(X) = - \sum_x p(x) \log p(x)$$

- ▶ Uniform the most uncertain. The more “peaky” the more certain.
- ▶ For continuous RV:

$$H(X) = - \int_x p(x) \log p(x) dx$$

