Probabilities for machine learning

Roland Memisevic

February 2, 2006
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 \textbf{random variable} and $p(\cdot)$ is its \textbf{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 $p(X = \text{value}) = 0$ 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.)
The interesting **properties** of RVs are usually just properties of their distributions (not surprisingly).

**Mean:**

\[ \mu = \sum x p(x) \]
The interesting **properties** of RVs are usually just properties of their distributions (not surprisingly).

Mean:

$$\mu = \sum_x p(x)x$$
The interesting **properties** of RVs are usually just properties of their distributions (not surprisingly).

**Mean:**

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

**Variance:**

$$\sigma^2 =$$
The interesting **properties** of RVs are usually just properties of their distributions (not surprisingly).

**Mean:**

\[ \mu = \sum_x p(x)x \]

**Variance:**

\[ \sigma^2 = \sum_x p(x)(x - \mu)^2 \]
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:** \( \sigma = \sqrt{\sigma^2} \)
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... \( p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \)
- Etc...
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

\[
\sum_{x,y} p(x, y) = 1, \quad p(x, y) \geq 0
\]
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:
    
    \[ p(y|x) = \frac{p(x, y)}{p(x)} \quad \text{and} \quad p(x|y) = \frac{p(x, y)}{p(y)} \]
  
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').
Independence and conditional independence

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

  \[ p(x, y) = p(x)p(y) \]

- 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

  \[ p(x, y|z) = p(x|z)p(y|z) \]
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 $K$.
- But what if all $x_i$ independent?
- Need to specify just $K$ probabilities, since the joint is the product!
- A more sophisticated version of this idea is to use *conditional* independence. Large and active area of 'Graphical Models'.
Maximum Likelihood

- Another useful thing about independence.
- Task: Given some data \((x_1, \ldots, x_N)\) build a \textit{model} of the data-generating process. Useful for classification, novelty detection, 'image manipulation', and countless other things.
- Possible solution: Fit a \textbf{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.)
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:

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

We set to zero and get:

\[
\mu = \frac{1}{N} \sum_i x_i
\]