PROBABILITY AND MATRIX DECOMPOSITION TUTORIAL

Paul Vicol

February 7, 2017

CSC 321, University of Toronto

TUTORIAL OUTLINE

1. Review of Probability
2. Expectation and Variance
3. Matrix Terminology (Symmetric, Positive Definite)
4. Eigendecomposition of Symmetric Matrices

WHY PROBABILITIES?

- A problem when building complex systems is **brittleness**
  - That is, when small irregularities cause models to break
- Probabilities are a great formalism for avoiding brittleness because they allow us to be explicit about uncertainties
- Instead of representing *values*, define *distributions over possibilities*

- Sum Rule (a.k.a Marginalization)
  - For discrete random variables:

$$p(X) = \sum_{Y} p(X, Y)$$

  - For continuous random variables:

$$p(X) = \int p(X, Y) dY$$

- Product Rule

$$p(X, Y) = p(Y|X)p(X)$$

- These two rules form the basis of all the complex probabilistic models we study

BAYES' RULE

- From the product rule, and symmetry, we have:

$$p(X, Y) = p(Y, X)$$

$$p(Y|X)p(X) = p(X|Y)p(Y)$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

- By the sum and product rules, the denominator is $p(X) = \sum_Y p(X, Y) = \sum_Y p(X|Y)p(Y)$
  - This is a *normalization constant* required to ensure that the sum of the conditional probabilities $p(Y|X)$ over all Y equals 1

- The *average value* of a function $f(x)$ under a probability distribution (or density) $p(x)$ is called the *expectation* of $f(x)$:

$$\mathbb{E}[f] = \sum_x p(x)f(x)$$

$$\mathbb{E}[f] = \int p(x)f(x)dx$$

- When we consider the expectation of a function of several variables, we use a subscript to indicate which variable is being averaged over:

$$\mathbb{E}_x[f(x,y)] = \sum_x p(x)f(x,y)$$

- Note that $\mathbb{E}_x[f(x,y)]$ is a function of y.
- Conditional expectation with respect to the conditional distribution:

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x)$$

APPROXIMATING THE EXPECTATION

- Given a finite number N of points *drawn from the probability distribution* $p(x)$, the expectation can be approximated as:

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

- This approximation becomes exact as $N \to \infty$

- The expectation is a *linear* operation:

$$\mathbb{E}[\mathsf{af}(\mathsf{x}) + \mathsf{bg}(\mathsf{x})] = \mathsf{a}\mathbb{E}[\mathsf{f}(\mathsf{x})] + \mathsf{b}\mathbb{E}[\mathsf{g}(\mathsf{x})]$$

$$\mathbb{E}[af(x) + bg(x)] = \sum_x p(x)[af(x) + bg(x)] =$$

$$a\sum_x p(x)f(x) + b\sum_x p(x)g(x) = a\mathbb{E}[f(x)] + b\mathbb{E}[g(x)]$$

- The variance of $f(x)$ measures how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$, and is defined by:

$$\text{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$$

- The variance can also be written in terms of the expectations of $f(x)$ and $f(x)^2$:

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

- Note that if $f(x) = x$ then:

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

- Everything we can possibly ask about a set of random variables $\{x_1, \ldots, x_n\}$ can be answered from the joint probability distribution $p(x_1, \ldots, x_n)$
- If we have many variables $x_1, x_2, \ldots, x_K$, then the joint distribution $p(x_1, \ldots, x_K)$ is huge, and intractable to deal with
- Two random variables x and y are **independent** iff

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

- x and y are **conditionally independent** given another random variable z iff

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

- The joint distribution can be *factored* into a product of simpler distributions by making *independence assumptions*
  - Probabilistic graphical models

TWO VIEWS OF PROBABILITY

- **ClassicalFrequentist interpretation**: views probabilities in terms of the frequencies of random, repeatable events.
- **Bayesian interpretation**: views probabilities as providing a quantification of uncertainty.
    - A more genral view
- The rules of probability arise naturally when numerical values are used to represent *degrees of belief*

MATRIX DECOMPOSITION

- An **eigenvector** of a square matrix $\mathbf{A}$ is a non-zero vector $\mathbf{v}$ such that

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

- The scalar λ is called the **eigenvalue** corresponding to the eigenvector $\mathbf{v}$
- A matrix $\mathbf{A}$ is *symmetric* iff

$$\mathbf{A} = \mathbf{A}^T$$

- A matrix $\mathbf{A}$ is *positive definite* iff for any vector $\mathbf{x}$:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$$

- We can gain insight about the properties of a matrix by decomposing it into constituent parts
- A square matrix A is said to be *diagonalizable* if there exists an invertible matrix P and a diagonal matrix D such that $A = PDP^{-1}$
- This is useful for finding *powers* of matrices
- If A is diagonalizable, then:

$$A^3 = (PDP^{-1})(PDP^{-1})(PDP^{-1}) = PD(P^{-1}P)D(P^{-1}P)DP^{-1} = PDDDP^{-1} = PD^3P^{-1}$$

- In general, if $A = PDP^{-1}$, then $A^k = PD^kP^{-1}$
- This is useful because it is easy to find powers of diagonal matrices:
- If $D = \begin{bmatrix} 7 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$, then $D^3 = \begin{bmatrix} 7^3 & 0 & 0 \\ 0 & (-2)^3 & 0 \\ 0 & 0 & 3^3 \end{bmatrix}$

- **Eigendecomposition** involves factorizing a matrix into a canonical form where it is represented in terms of its **eigenvectors** and **eigenvalues**
- Given a matrix A that has n linearly independent eigenvectors, A can be factored as:

$$A = Q\Lambda Q^{-1}$$

  - Q is a matrix whose columns are the eigenvectors of A
  - Λ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues of A

- When A is *symmetric*, its eigenvectors can be chosen to be orthogonal, so we have:

$$A = Q\Lambda Q^{T}$$

REFERENCES AND RESOURCES

- Deep Learning Book - Eigendecomposition
  http://www.deeplearningbook.org/contents/linear_algebra.html
- Matrix Calculus Reference
  http://www.atmos.washington.edu/~dennis/MatrixCalculus.pdf
- Pattern Recognition and Machine Learning (Book), by Christopher Bishop