Lab 1: Inference and Representation

Rahul G. Krishnan

New York University

September 7, 2016

Welcome!

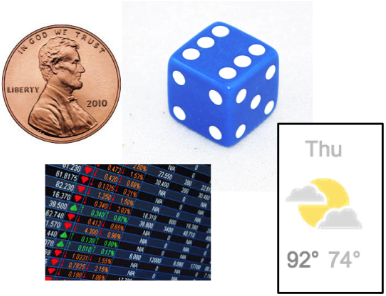
- Instructor: David Sontag (dsontag [at] cs.nyu.edu) and Joan Bruna (bruna [at] cs.nyu.edu)
- Lab Instructor: Rahul G. Krishnan, (rahul [at] cs.nyu.edu)
- Lecture: Mon: 5-7pm, Warren Weaver Hall 1302
- Lab: Wed 7-8pm, Meyer Hall of Physics 121
- Website: `https://github.com/inf16nyu/home`
- Forum: Piazza (Enroll in it! You should have gotten an email with a link to join the class already.)
- Attendance required for both lab and lecture (may be taken periodically)
- Lab material will be complementary to main lectures and will be posted on website

Overview of some basics topics you will encounter

- Probability Review: Random variables, probability distributions, marginalization, Bayes Rule
- Optimization Review: Convexity, maximization, minimization, Jensen's Inequality

Random Variables

- Examples:



- Discrete or continuous

Probability Distributions

- All random variables have a (possibly unknown) probability distribution
- Lets call Ω the sample space and x a state taken by the random variable in Ω
- Intuition: The distribution maps x to the likelihood of the state (a positive real number)
- Formally: p: $\Omega \to \mathbb{R}$ and $p(x) >= 0$ for all $x \in \Omega$
- Probability density must sum to 1:

$$\underbrace{\sum_{x \in \Omega} p(x) = 1}_{\text{Discrete RVs}} \qquad \underbrace{\int_{x \in \Omega} p(x)dx = 1}_{\text{Continuous RVs}}$$

Examples of Probability Distributions

IPython Notebook : Random Variables

Expectation

- The expectation of the random variable under its probability distribution is called the mean, average or first moment
- $$\underbrace{\mathbb{E}(X) = \sum_x xp(x)}_{\text{Discrete}} \qquad \underbrace{\mathbb{E}(X) = \int_x xp(x)dx}_{\text{Continuous}}$$
- You can define expectations over any function of X, eg. Shannon Entropy (which measures the uncertainty in a random variable) is defined as $\mathbb{E}[-\log p(x)]$.
- Linearity of expectation:

$$\mathbf{E}(aX + b) = a\mathbf{E}(X) + b$$

Multivariate Distributions

- A single random variable isn't interesting, modeling, multiple random variables is more realistic
- Example: X = "sprinkler on", Y = "raining"

p(X,Y)	sprinkler off	sprinkler on
not raining	.6	.24
raining	.15	.01

- This is an example of a *multivariate* or *joint distribution*, i.e. $p(X, Y)$ or more generally $p(X_1, X_2, \ldots, X_n)$
- For discrete variables, defined as

$$p(X_1 = x_1, \ldots, X_n = x_n) :=$$
$$p(X_1 = x_1 \cap \ldots \cap X_n = x_n)$$

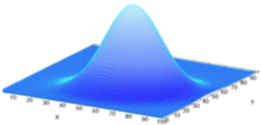
The Multivariate Gaussian Distribution

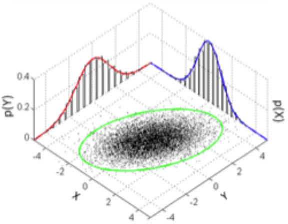
- One variable, with parameters μ, σ

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

- Multiple variables (in vector **x**), with parameters $\boldsymbol{\mu}, \Sigma$

$$p(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^T\Sigma^{-1}(x-\boldsymbol{\mu})\right)$$





Marginalization

- Suppose we have joint distribution $p(X_1, \ldots, X_n)$ and we want to know $p(X_i = x_i)$
- Example: from p(X,Y), what is $p(Y = \text{raining})$?

p(X,Y)	sprinkler off	sprinkler on
not raining	.6	.24
raining	.15	.01

- Answer: .15 + .01 = .16
- More generally:

$$p(X_i = x_i) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \sum_{x_n} p(x_1, \ldots, x_n)$$

- Doing this for each x_i gives us the *marginal distribution* over X_i

Conditioning

- What if we want to know the distribution of X when Y is set to a particular value?
- Example: from p(X,Y), what is *p*(sprinkler on | raining)?

p(X,Y)	sprinkler off	sprinkler on
not raining	.6	.24
raining	.15	.01

- Answer: .01 / (.15 + .01) = 1/16
- More generally:

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

- Note that this is a univariate distribution, defined for each value of y

Independence

- Intuitively: two variables are independent if they are unrelated
- One definition: $X \perp Y$ means that for all x and y, $p(x, y) = p(x)p(y)$
- Examples: The identity (head/tails) of two different tosses of the same coin

Chain Rule

Chain Rule:

$$p(x_1,\ldots,x_n) = \underbrace{\underbrace{p(x_1)p(x_2|x_1)}_{p(x_1,x_2)}p(x_3|x_1,x_2)}_{p(x_1,x_2,x_3)\ldots}\ldots p(x_n|x_1,x_2,\ldots,x_{n-1})$$

- Specifies how you can write the joint probability as the product of conditional probabilities

Bayes' Rule

Bayes' Rule:

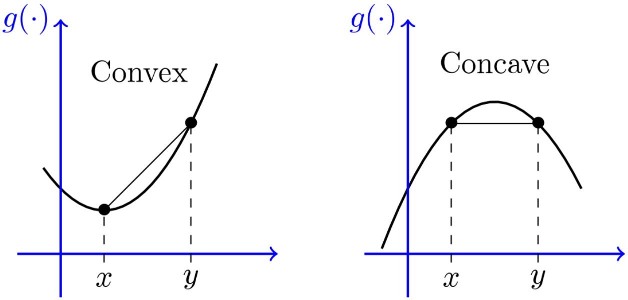
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

- Basis of Bayesian statistics
- Useful when $p(y|x)$ is more natural to measure or estimate than $p(x|y)$
- Proof: From chain rule and definition of conditional distribution

Overview

- Functions can vary from simple lines (linear) to complicated curves (non-linear)
- We're often interested in finding points that achieve maxima, minima (local or global) in the function
- There are a few types of functions that are interesting and worth knowing

Convexity

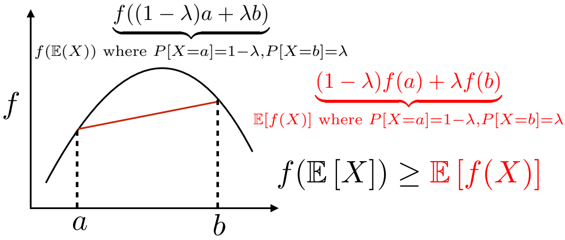


- Convex functions have unique global minima
- Concave functions unique global maxima
- For a convex (or concave) function, starting from any point along the surface, and following the gradient (or its negative) yields a point where the function reaches its minima (or maxima) (conditions apply :))

Why is this relevant?

- Probability densities are functions (albeit complicated ones) in the parameters of the distribution. eg: $p(x; \mu, \Sigma)$ is a function of x, μ, Σ.
- Probablistic inference is often concerned with fitting parameters. i.e finding μ, Σ such that x is maximially likely under the resulting probability distribution.
- This will be a theme moving forward in the class

Jensen's Inequality



For *f* concave:

- It is an important identity to know and understand because it shows up in probablistic inference
- Lookahead: We will use this identity when we're interested in learning latent variable models

Questions?

- Please read probability review: http://cs229.stanford.edu/section/cs229-prob.pdf

Acknowledgements: Rachel Hodos, whose slides I built off of.