

CS281A/Stat241A recitation 1: Review

August 30, 2007

Percy Liang

Probability

- Elementary (20 minutes)
 - Relevance: foundation
 - Joint, marginal, conditional probabilities
 - * Joint distribution specifies everything, use laws of probability to deduce other quantities
 - * Important computation: sums/integrals; intuition is to incorporate information about related variables that we're not interested in
 - * Chain rule
 - Bayes rule: flip conditional probabilities; simple derivation
 - Useful in Bayesian statistics to compute $p(\theta | x)$; replace y with θ
 - Note: can use Bayes rule when not being Bayesian
 - * When write $p(x | \theta)$, is it specified or derived? Specify local factors, form joint, then deduce other quantities
 - Independence and conditional independence
 - * Key property that makes probability a rich theory; nice relationship to computation
 - * $p(x)p(y) = p(x, y)$ for all x, y
 - * $p(x) = p(x | y)$ for all x, y
 - * CI: need to be independent for all values of the conditioning variable
 - * Notation: $Y \perp Z | X$
 - * General: all statements implicitly have for all values of x, y
- Formalities (5 minutes)
 - Relevance: be more formal (none in class)
 - Random variables X versus outcomes $\omega \in \Omega$; layer of abstraction allows to fix the model and talk about different quantities: $X^2, \mathbb{1}[X = a]$
 - Notation: random variable (uppercase) versus value (lowercase),
 - Discrete versus continuous variables (don't make a big deal of)

- Basic family of distributions (10 minutes)
 - Relevance: build basic modules
 - Notation, data types: random variables X , values x , distributions \mathcal{N} , probability function $p(x)$
 - Binomial/multinomial: $X \sim \text{Binomial}(n, \theta)$
 - Gaussian: $X \sim \mathcal{N}(\mu, \sigma^2)$
 - (Mention quickly: Beta, Poisson, Gamma, Beta)
- Expectation, moments (20 minutes)
 - Relevance: computations with distributions
 - (Point: get at general properties of the distribution)
 - Definition: $\mathbb{E}f(X) = \sum_x p(X = x)f(x) = \int_x f(x)p(x)dx$
 - Mean: $\mathbb{E}X$ (e.g., Gaussian)
 - Variance: $\text{var}(X) = \mathbb{E}[X - \mathbb{E}X]^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$ (e.g., Gaussian)
 - Indicator: $\mathbb{E}\mathbb{1}[X = x] = p(X = x)$ (e.g., multinomial; to get back probabilities)
 - Covariance: $\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$
 - * Definition: if $\text{cov}(X, Y) = 0$, X and Y are uncorrelated
 - Properties:
 - * Linearity: $\mathbb{E}[aX] = a\mathbb{E}X$, $\mathbb{E}[X + Y] = \mathbb{E}X + \mathbb{E}Y$
 - * If uncorrelated, then can separate: $\mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y$
 - Uncorrelated is not independent (diamond example)
 - General: remember what's a constant and what's a random variable; for example, $\mathbb{E}[X\mathbb{E}Y] = \mathbb{E}X\mathbb{E}Y$
 - Comment: moments are different from parameters in general; consider Beta
 - Parameterization:
 - * Replace variance σ^2 with precision $\rho = 1/\sigma^2$
 - Comment: can parameterize distributions using one-to-one transformation
- Convexity, Jensen's inequality (5 minutes)
 - Relevance: EM, variational methods
 - Definition: $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$
 - Follows: $\leq f(\mathbb{E}X) \leq \mathbb{E}f(X)$
 - Sufficient condition: if $f''(x) > 0$, then convex
 - Example function (important): $f(x) = \log x$ is concave

- Limit theorems (5 minutes)
 - Relevance: limit of infinite data (not in this class)
 - Law of large numbers:
 - * X_i i.i.d. (assume integrable)
 - * $\sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}X$
 - Central limit theorem
 - * X_i i.i.d. (assume finite variance)
 - * $\sqrt{n}(\sum_{i=1}^n (X_i - \mathbb{E}X)) \xrightarrow{d} \mathcal{N}(0, \text{var}(X))$
 - (Mention quickly: Large deviation theory)

Statistics

- Maximum likelihood: moments \leftarrow parameters \rightarrow data
 - Definition: $\theta_{\text{ML}} = \text{argmax}_{\theta} \log p(x; \theta)$
 - Why log?
 - * Monotonic: preserve argmax
 - * Makes differentiation easier:
 - Turns products into sums
 - Turns exponentiation into product
 - General: $\sum \log$ is good, $\log \sum$ is bad
 - Main principle that we'll use in this class (others are possible)
 - Can do in closed form (in general, no closed form)
 - * Derivation for Gaussian mean and variance
 - * Assume second derivative is convex
 - * Homework: derivation for binomial
 - Moments not same as parameters (again)
 - Mention quickly:
 - * Exponential families \rightarrow existence/uniqueness
 - * Mixture models, hard (Jensen's inequality to push summation inside)

Linear algebra

- Relevance: build more advanced modules (PCA, regression, Kalman filters), kernel methods, convex optimization
- Motivation: interpret multivariate Gaussian

- Definition: covariance matrix of a random vector; $\text{var}(X) = \mathbb{E}(X - \mu)(X - \mu)^T$
 (i, j) -th entry is $\mathbb{E}(X_i - \mu)(X_j - \mu)$
- Symmetric positive semi-definite: $A \succeq 0$,
 - * Definition: $b^T A b \geq 0$ for all $b \neq 0$
 - * Eigenvalues are all nonnegative
 - Eigenvectors, eigenvalues: $Av = \lambda v$
 - Eigen decomposition: $A = U \Lambda U^T = \sum_i \lambda_i u u^T$
 - * Is the covariance matrix of some random vector
 - Show that if $A = R^T R$, then $A \succeq 0$,
 - Aside: every covariance matrix is positive semi-definite
 - $A = (\Lambda^{1/2} U^T)^T (\Lambda^{1/2} U^T)$
 - * Sylvester's condition (for positive definiteness not semi-definite): $|A_k| > 0$ for all $k = 1, \dots, d$
 - Definition: determinant $|A| = 0$
 - Properties: $|AB| = |A||B|$, $|A^{-1}| = |A|^{-1}$
 - Interpretation: Product of eigenvalues (volume of Gaussian)
- Show: covariance matrix always positive semi-definite
 - * $\text{var}(b^T X) = \mathbb{E} b^T (X X^T) b = b^T \text{var}(X) b$
- Not cover: general
 - Matrix decomposition: SVD
 - Trace, inverse
- Vector calculus (least squares)
- For regression (geometric intuition)
 - Projection: dot products

Graph theory

- Relevance: get at properties of distributions
- Directed (we only work with DAGs) versus undirected
- Cycles: what it means in DAGs, undirected
- Trees: linear number of edges
- Minimum spanning tree algorithm

Information theory

- Entropy, conditional entropy
- KL-divergence
- Mutual information