CS4801 : Recap of Probability

Sahely Bhadra
10/8/2017

1. Probability basic
2. Maximum Likelihood Estimation

Probability Basic

Random Variables and Densities

- Random variables X represents outcomes or states of world.
  Instantiations of variables usually in lower case: x
  We will write $p(x)$ to mean probability($X = x$).
- Sample Space: the space of all possible outcomes/states.
  (May be discrete or continuous or mixed.)
- Probability mass (density) function $p(x) \geq 0$
  Assigns a non-negative number to each point in sample space.
  Sums (integrates) to unity: $\sum_x p(x) = 1$ or $\int_x p(x)dx = 1$.
  Intuitively: how often does x occur, how much do we believe in x.
- Ensemble: random variable $+$ sample space$+$ probability function

Probability Basic

Expectations, Moments

- Expectation of a function $a(x)$ is written $E[a]$ or $\langle a \rangle$

$$E[a] = \langle a \rangle = \sum_x p(x)a(x)$$

  e.g. mean $= \sum_x xp(x)$, variance $= \sum_x (x - E[x])^2 p(x)$

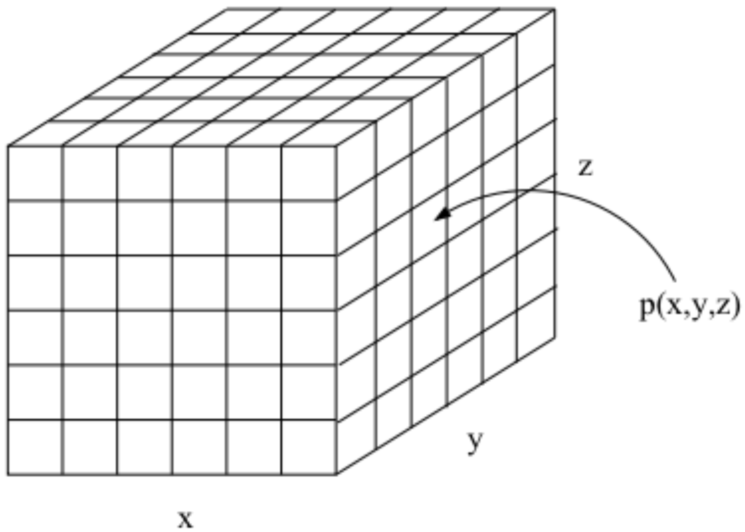
- Moments are expectations of higher order powers.
  (Mean is first moment. Autocorrelation is second moment.)
- Centralized moments have lower moments subtracted away (e.g. variance, skew, curtosis).
- Deep fact: Knowledge of all orders of moments completely defines the entire distribution.

Probability Basic

Joint Probability

- Key concept: two or more random variables may interact. Thus, the probability of one taking on a certain value depends on which value(s) the others are taking.
- We call this a joint ensemble and write

$$p(x, y) = \text{prob}(X = x \text{ and } Y = y)$$

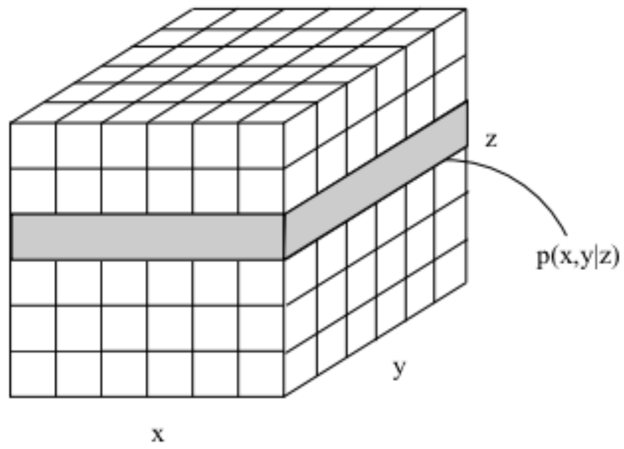


Probability Basic

Conditional Probability

- If we know that some event has occurred, it changes our belief about the probability of other events.
- This is like taking a "slice" through the joint table.

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



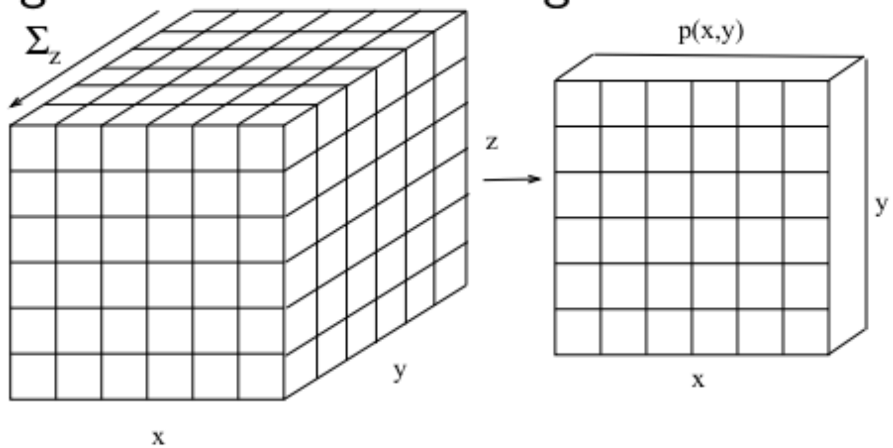
Probability Basic

Marginal Probabilities

- We can "sum out" part of a joint distribution to get the *marginal distribution* of a subset of variables:

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

- This is like adding slices of the table together.



- Another equivalent definition: $p(x) = \sum_y p(x|y)p(y)$.

Probability Basic

Bayes' Rule

- Manipulating the basic definition of conditional probability gives one of the most important formulas in probability theory:

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

- This gives us a way of "reversing" conditional probabilities.
- Thus, all joint probabilities can be factored by selecting an ordering for the random variables and using the "chain rule":

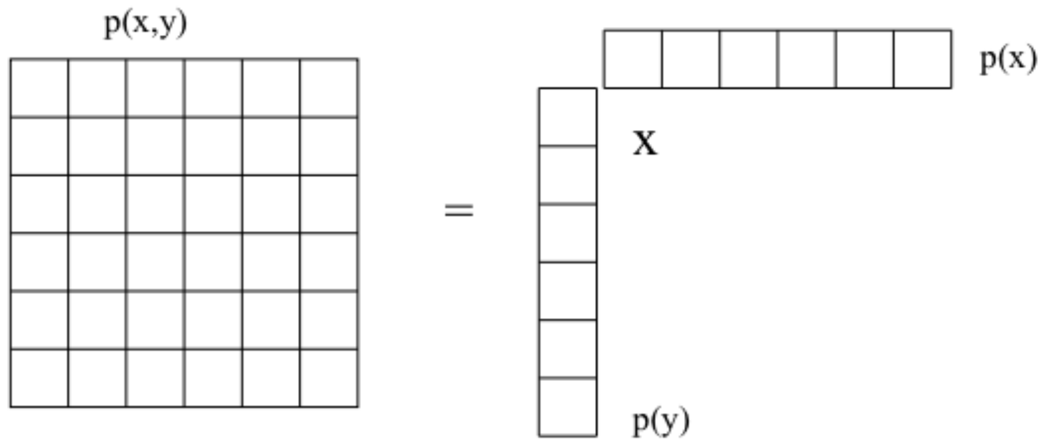
$$\boxed{p(x, y, z, \ldots) = p(x)p(y|x)p(z|x, y)p(\ldots|x, y, z)}$$

Probability Basic

Independence & Conditional Independence

- Two variables are independent iff their joint factors:

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



- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

$$p(x, y|z) = p(x|z)p(y|z) \qquad \forall z$$

Probability Basic

Bernoulli

- For a binary random variable with p(heads)$=\pi$:

$$p(x|\pi) = \pi^x (1-\pi)^{1-x}$$

Multinomial

- For a set of integer counts on k trials

$$p(\mathbf{x}|\pi) = \frac{k!}{x_1! x_2! \cdots x_n!} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_n^{x_n} = h(\mathbf{x}) \exp\left\{ \sum_i x_i \log \pi_i \right\}$$

- But the parameters are constrained: $\sum_i \pi_i = 1$.

Probability Basic

Gaussian (normal)

- For a continuous univariate random variable:

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

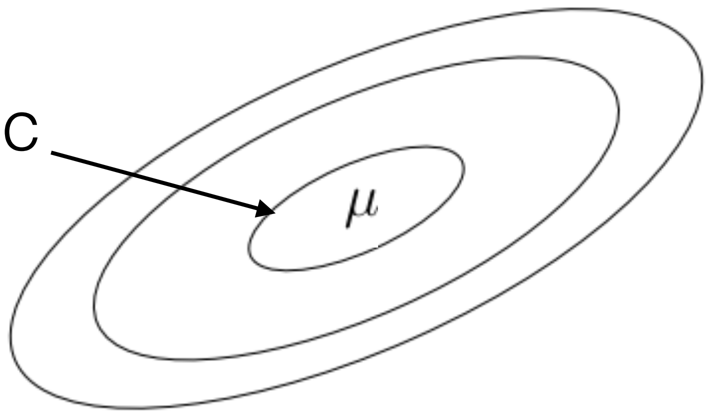
Multivariate Gaussian Distribution

- For a continuous vector random variable:

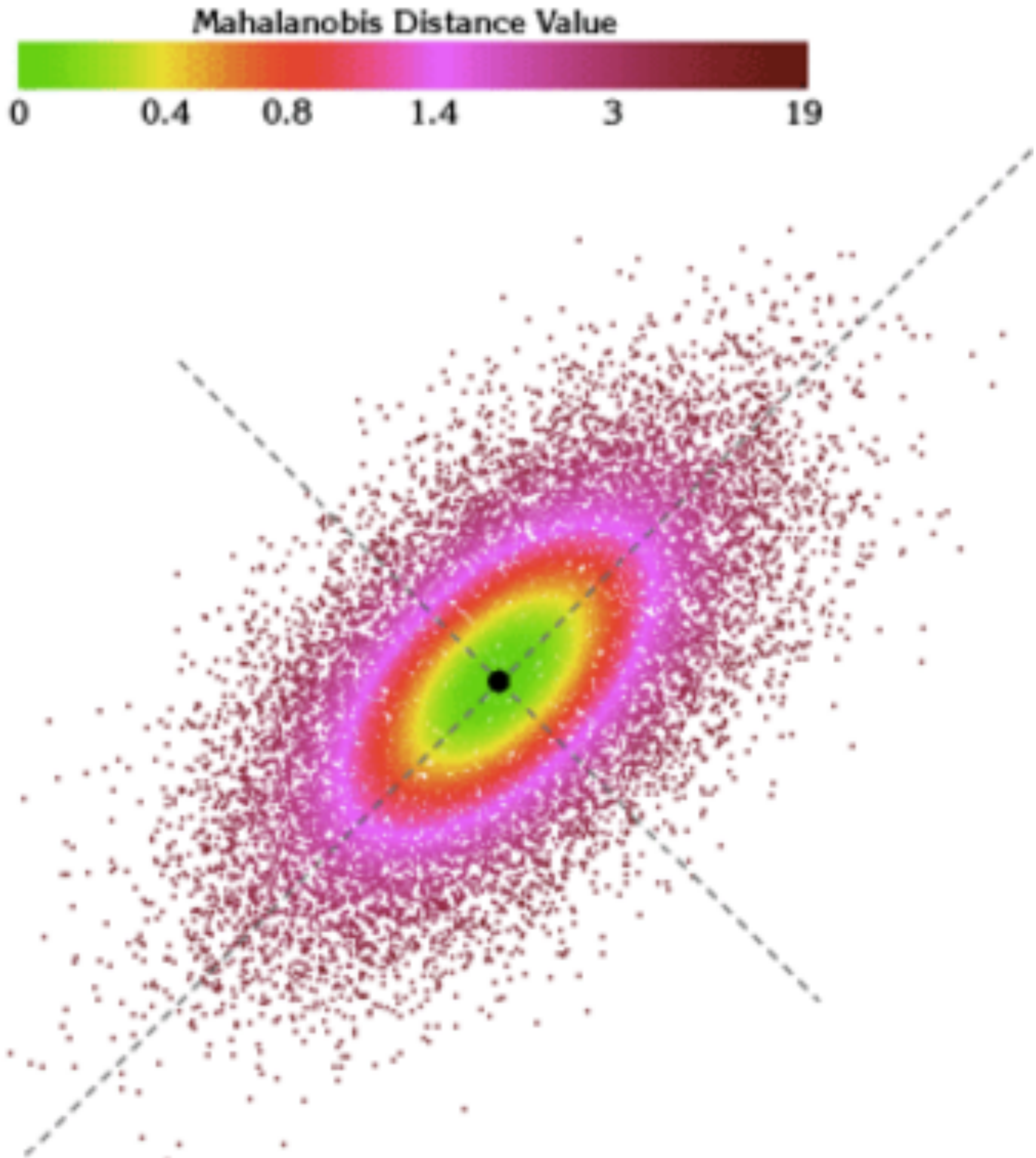
$$p(x|\mu,\Sigma) = |2\pi\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

$$\underline{(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)} = C$$

Mahalanobis distance



Probability Basic



Maximum Likelihood Estimation(MLE)

Bernoulli distribution

$$p(D / \theta) = \theta^{x_1}(1-\theta)^{(1-x_1)} \cdots \theta^{x_n}(1-\theta)^{(1-x_n)} = \theta^{(x_1+\cdots+x_n)}(1-\theta)^{n-(x_1+\cdots+x_n)}.$$

Log likelihood function

$$\ln p(D / \theta) = \ln\theta(\sum_{i=1}^{n} x_i) + \ln(1-\theta)(n - \sum_{i=1}^{n} x_i) = n\bar{x}\ln\theta + n(1-\bar{x})\ln(1-\theta).$$

MLE

$$\hat{\theta}(\mathbf{x}) = \bar{x}.$$

Maximum Likelihood Estimation(MLE)

Gaussian distribution

p(D/ μ, σ^2)

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_1-\mu)^2}{2\sigma^2}\right) \cdots \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_n-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2.$$

log likelihood function

$$\ln p(D/\mu, \sigma^2) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2.$$

$$\frac{\partial}{\partial\mu} \ln p(D/\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i-\mu) = \frac{1}{\sigma^2} n(\bar{x}-\mu)$$

$$\frac{\partial}{\partial\sigma^2} \ln p(D/\mu, \sigma^2) = -\frac{n}{\sigma^2} + \frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (x_i-\mu)^2 = \frac{n}{(\sigma^2)^2} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} (x_i-\mu)^2\right).$$

MLE

$$\hat{\mu}(\mathbf{x}) = \bar{x}$$

$$\hat{\sigma}^2(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x})^2.$$

Next Classes

- 14/8
  - Probabilistic interpretation of Ridge regression.

- 15/8 : Do you want an alternative class ?