Review of Probability

ECE57000: Artificial Intelligence

David I. Inouye

Sep. 18, 2019

Announcements

- Quiz 3 on Monday (Sep. 23)

- Updated syllabus
  - No longer require scikit-learn interface
  - Can use any Python libraries (e.g., PyTorch, TensorFlow or Keras)

- Paper selection grades are out

Note: Conditional and marginal distributions can be computed for *any set of variables*

- Suppose $p(\boldsymbol{x}) = p(x_1, x_2, x_3, x_4)$

$$p(x_1, x_3) = \int_{x_2, x_4} p(\boldsymbol{x}) dx_2 dx_4$$

$$p(x_1, x_2 | x_3) = \frac{p(x_1, x_2, x_3)}{p(x_3)}$$

$$= \frac{\int_{x_4} p(\boldsymbol{x}) dx_4}{\int_{x_1, x_2, x_4} p(\boldsymbol{x}) dx_1 dx_2 dx_4}$$

Chain rule (or product rule) of probability

- The joint distribution can be written as product of conditional PDFs/PMFs:

$$p(x_1, x_2) = p(x_1)p(x_2|x_1)$$

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)$$

- This can be written as:

$$p(x_1, x_2, \dots, x_d) = \prod_{i=1}^{d} p(x_i|x_1, \dots, x_{i-1})$$

- Consequence (order doesn't matter):

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

<u>Bayes rule</u>: Enables conversion between one conditional and the other (they are ***different***)

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

(derive on board, show reverse)

When are $p(x|y)$ and $p(y|x)$ equal?

Independence means that one variable is not affected by the other variable

- Example: Flip two coins, X and Y are 0 or 1.
- Counterexample: Roll dice for number X; then flip that number of coins and count the number of heads Y.
- Formally, PDF/PMF can be written as product of functions that only involve x or y (but not both)

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

- Usually, these are the marginal densities:

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

- Equivalent definition:

$$p(x|y) = p(x) \text{ and } p(y|x) = p(y)$$

Two variables are <u>conditionally independent</u> if they are independent conditioned on a third variable

- Example: Person A is home late (event X), Person B is home late (event Y), snowstorm hits West Lafayette (event Z)
- Formally, X and Y are conditionally independent given Z if:
$$p(x, y, z) = f(x, z) f(y, z)$$
$$p(x, y|z) = p(x|z) p(y|z)$$
- Notation: Independence $X \perp Y$
- Notation: Conditional independence $X \perp Y \mid Z$

An <u>expectation</u> (or <u>expected value</u>) of a function of a random variable is the average or mean value with respect to its distribution

- Formal definitions

$$\mathbb{E}_{X\sim P(x)}[f(x)] \equiv \sum_{x\in X} f(x)P(x)$$

$$\mathbb{E}_{X\sim p(x)}[f(x)] \equiv \int_{x\in X} f(x)p(x)dx$$

- Sometimes drop notation to $\mathbb{E}_X[f(x)]$ or just $\mathbb{E}[f(x)]$ if clear from context
- Common: Mean of the distribution $\mu = \mathbb{E}[x]$
- Examples: $P(x) = [0.4,0.3,0.1,0.3]$, $p(x) = 3x^2$

Expectation is a *linear operator* (i.e. splits on summation and scale can come out)

- A linear operator H must satisfy two properties:

$$H\big(f(x) + g(x)\big) = H\big(f(x)\big) + H\big(f(y)\big)$$

$$H\big(\alpha f(x)\big) = \alpha H\big(f(x)\big)$$

- Derive for matrix operator and vector

- Derive for expectations, i.e. when $H = \mathbb{E}$

Variance measures the "spread" of a distribution

- Definition

$$\mathrm{Var}[x] = \sigma^2 \equiv \mathbb{E}_X[(x-\mu)^2]$$

$$= \mathbb{E}_X[(x - \mathbb{E}_X[x])^2]$$

- Intuitively, recenter and then measure expected value of $f(x) = x^2$

- **Standard deviation** is square root of variance

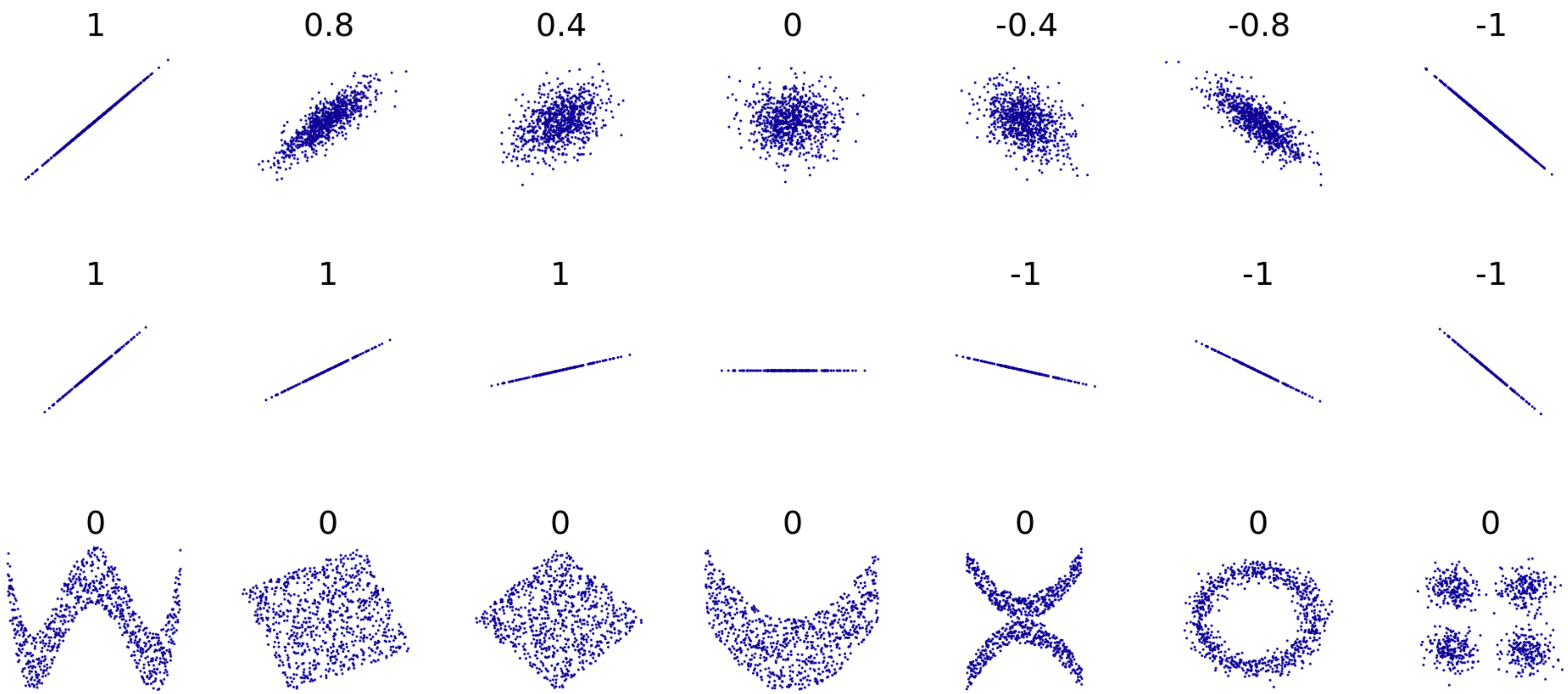
$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathbb{E}_X[(x-\mu)^2]}$$

Covariance and correlation measure *linear* relationship between two variables

- Covariance definition
$$\mathrm{Cov}[x,y] \equiv \sigma_{X,Y}^2 \equiv \mathbb{E}_{X,Y}\big[(x-\mu_X)\big(y-\mu_y\big)\big]$$

- Correlation is a normalized covariance
$$\rho_{X,Y} \equiv \frac{\sigma_{X,Y}^2}{\sigma_X \sigma_Y}$$
- Example: $P(x,y) = \big[[0.4, 0.1], [0.1, 0.4]\big]$
- $\mu_X = \mu_Y = 0.5, \sigma_X^2 = \sigma_Y^2 = 0.25$
- $\sigma_{X,Y}^2 = -\frac{3}{20}, \rho_{X,Y} = -\frac{3}{5}$

Uncorrelated ($\rho_{X,Y} = 0$) is **NOT** the same as independence (because only measures *linear* relationship)



Covariance and correlation matrix

are generalizations for vectors

- Covariance matrix has covariance of every pair of random variables

$$\Sigma = \begin{bmatrix} \sigma^2_{X_1,X_1} & \cdots & \sigma^2_{X_1,X_d} \\ \vdots & \ddots & \vdots \\ \sigma^2_{X_d,X_1} & \cdots & \sigma^2_{X_d,X_d} \end{bmatrix}$$

- Matrix has variance along diagonal $\sigma^2_{X_i,X_i} = \sigma^2_{X_i}$
- Correlation matrix is similar but with 1s on diagonal

$$\mathrm{R} = \begin{bmatrix} 1 & \cdots & \rho_{X_1,X_d} \\ \vdots & \ddots & \vdots \\ \rho_{X_d,X_1} & \cdots & 1 \end{bmatrix}$$

- Both matrices are symmetric $\Sigma = \Sigma^T$ and $R = R^T$

The *empirical* distribution and *empirical* expectation are *sampled* versions of their counterparts

- Dirac delta function is a point mass at μ
$$\delta(x-\mu) \equiv \lim_{\sigma^2 \to 0^+} \mathcal{N}(x; \mu, \sigma^2)$$
- **Empirical distribution** is formed from samples $\{x_i\}_{i=1}^n$
$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$
- **Empirical expectation** is expectation with respect to the empirical distribution (i.e., average over samples)
$$\widehat{\mathbb{E}}[f(x)] = \int_x f(x) \hat{p}(x) dx = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

Informally, **entropy** measures the “amount of randomness/disorder” of a distribution

- Formally, **entropy** for discrete variables

$$H\big(P(x)\big) = \mathbb{E}[-\log P(x)] = \sum_x -P(x)\log P(x)$$

- Formally, **differential entropy** for continuous variables

$$H\big(p(x)\big) = \mathbb{E}[-\log p(x)] = \int_x -p(x)\log p(x)\,dx$$

- Consider fair coin vs coin where both sides are heads