Probabilistic Machine Learning (2): Probability Basics (Contd)

CS771: Introduction to Machine Learning

Piyush Rai

Expectation

- Expectation of a random variable tells the expected or average value it takes

- Expectation of a discrete random variable $X \in S_X$ having PMF $p(X)$

$$\mathbb{E}[X] = \sum_{x \in S_X} x p(x)$$

Probability that $X = x$

- Expectation of a continuous random variable $X \in S_X$ having PDF $p(X)$

$$\mathbb{E}[X] = \int_{x \in S_X} x p(x) dx$$

Probability density at $X = x$

Note that this exp. is w.r.t. the distribution $p(f(X))$ of the r.v. $f(X)$

- The definition applies to functions of r.v. too (e.g.., $\mathbb{E}[f(X)]$)

Often the subscript is omitted but do keep in mind the underlying distribution

- Exp. is always w.r.t. the prob. dist. $p(X)$ of the r.v. and often written as $\mathbb{E}_p[X]$

Expectation: A Few Rules

X and Y need not be even independent. Can be discrete or continuous

- Expectation of sum of two r.v.'s: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- Proof is as follows
  - Define $Z = X + Y$

$$\mathbb{E}[Z] = \sum_{z \in S_Z} z \cdot p(Z = z) \qquad \text{s.t. } z = x + y \text{ where } x \in S_X \text{ and } y \in S_Y$$

$$= \sum_{x \in S_X} \sum_{y \in S_Y} (x + y) \cdot p(X = x, Y = y)$$

$$= \sum_x \sum_y x \cdot p(X = x, Y = y) + \sum_x \sum_y y \cdot p(X = x, Y = y)$$

$$= \sum_x x \sum_y p(X = x, Y = y) + \sum_y y \sum_x p(X = x, Y = y)$$

$$= \sum_x x \cdot p(X = x) + \sum_y y \cdot p(Y = y)$$

Used the rule of marginalization of joint dist. of two r.v.'s

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

Expectation: A Few Rules (Contd)

α is a real-valued scalar

α and β are real-valued scalars

f and g are arbitrary functions.

- Expectation of a scaled r.v.: $\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$
- Linearity of expectation: $\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$
- (More General) Lin. of exp.: $\mathbb{E}[\alpha f(X) + \beta g(Y)] = \alpha \mathbb{E}[f(X)] + \beta \mathbb{E}[g(Y)]$
- Exp. of product of two independent r.v.'s: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Law of the Unconscious Statistician (LOTUS): Given an r.v. X with a known prob. dist. $p(X)$ and another random variable $Y = g(X)$ for some function g

Requires finding $p(Y)$

Requires only $p(X)$ which we already have

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_{y \in S_Y} y p(y) \quad = \sum_{x \in S_X} g(x) p(x)$$

LOTUS also applicable for continuous r.v.'s

- Rule of iterated expectation: $\mathbb{E}_{p(X)}[X] = \mathbb{E}_{p(Y)}[\mathbb{E}_{p(X|Y)}[X|Y]]$

Variance and Covariance

- Variance of a scalar r.v. tells us about its spread around its mean value $\mathbb{E}[X] = \mu$

$$\mathrm{var}[X] = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2$$

- Standard deviation is simply the square root is variance
- For two scalar r.v.'s X and Y, the covariance is defined by

$$\mathrm{cov}[X,Y] = \mathbb{E}[\{X-\mathbb{E}[X]\}\{Y-\mathbb{E}[Y]\}] = \mathbb{E}[XY] - \mathbb{E}[\mathrm{X}]\mathbb{E}[\mathrm{Y}]$$

- For two vector r.v.'s X and Y (assume column vec), the covariance matrix is defined by

$$\mathrm{cov}[X,Y] = \mathbb{E}[\{X-\mathbb{E}[X]\}\{Y^\top-\mathbb{E}[Y^\top]\}] = \mathbb{E}[XY^\top] - \mathbb{E}[\mathrm{X}]\mathbb{E}[Y^\top]$$

- Cov. of components of a vector r.v. X: $\mathrm{cov}[X] = \mathrm{cov}[X,X]$
- Note: The definitions apply to functions of r.v. too (e.g., $\mathrm{var}[f(X)]$)
- Note: Variance of sum of independent r.v.'s: $\mathrm{var}[X+Y] = \mathrm{var}[X] + \mathrm{var}[Y]$

Important result

Transformation of Random Variables

- Suppose $Y = f(X) = AX + b$ be a linear function of a vector-valued r.v. X (A is a matrix and b is a vector, both constants)
- Suppose $\mathbb{E}[X] = \mu$ and $\mathbf{cov}[X] = \Sigma$, then for the vector-valued r.v. Y

$$\mathbb{E}[Y] = \mathbb{E}[AX + b] = A\mu + b$$

$$\text{cov}[Y] = \text{cov}[AX + b] = A\Sigma A^\top$$

- Likewise, if $Y = f(X) = a^\top X + b$ be a linear function of a vector-valued r.v. X (a is a vector and b is a scalar, both constants)
- Suppose $\mathbb{E}[X] = \mu$ and $\mathbf{cov}[X] = \Sigma$, then for the scalar-valued r.v. Y

$$\mathbb{E}[Y] = \mathbb{E}[a^\top X + b] = a^\top \mu + b$$

$$\text{var}[Y] = \text{var}[a^\top X + b] = a^\top \Sigma a$$

Common Probability Distributions

Important: We will use these extensively to model data as well as parameters of models

- Some common discrete distributions and what they can model
  - **Bernoulli:** Binary numbers, e.g., outcome (head/tail, 0/1) of a coin toss
  - **Binomial:** Bounded non-negative integers, e.g., # of heads in n coin tosses
  - **Multinomial/multinoulli:** One of K (>2) possibilities, e.g., outcome of a dice roll
  - **Poisson:** Non-negative integers, e.g., # of words in a document
- Some common continuous distributions and what they can model
  - **Uniform:** numbers defined over a fixed range
  - **Beta:** numbers between 0 and 1, e.g., probability of head for a biased coin
  - **Gamma:** Positive unbounded real numbers
  - **Dirichlet:** vectors that sum of 1 (fraction of data points in different clusters)
  - **Gaussian:** real-valued numbers or real-valued vectors

Coming up next

- Probabilistic Modeling
- Basics of parameter estimation for probabilistic models