MATH5836: Data and Machine Learning

Week 0: Basics of Probability Theory

Sarat Moka *UNSW, Sydney*

Key Topics

0.3.1 Random Variables 0-2
0.3.2 Divergences and Entropies 0-4
0.3.3 Computations for Multivariate Normal Distributions 0-5

Reference:

- *Introduction to Probability* by Joseph K. Blitzstein & Jessica Hwang [click here for a pdf copy]
- Appendix B of *Mathematical Engineering of Deep Learning* by Liquet, Moka, and Nazarathy: Freely available at `https://deeplearningmath.org/`

0.3.1 Random Variables

Set Theory Basics

- **Sample Space (Ω)**: The set of all possible outcomes of an experiment.
- **Event**: A subset of Ω (e.g., $A \subseteq \Omega$).
- **σ-Algebra ($\mathcal{F}$)**: A collection of events closed under complements, countable unions, and intersections.
- **Random Variable**: A function $X : \Omega \to \mathbb{R}$ is a random variable if $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.

Discrete Random Variables

- **Definition**: X takes countable values (e.g., integers), denote them by $\mathscr{X} \subset \mathbb{R}$.
- **Probability Mass Function (PMF)**: $p_X(x) = P(X = x)$ for $x \in \mathscr{X}$.
- **Examples**:
  - Bernoulli: $p_X(1) = p$, $p_X(0) = 1 - p$.
  - Binomial: $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Continuous Random Variables

- **Definition**: X takes uncountably infinite values (e.g., real numbers).
- **Probability Density Function (PDF)**: $f_X(x)$ satisfies

$$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$$

- **Examples**:
  - Uniform: $f_X(x) = \frac{1}{b-a}$ for $x \in [a, b]$.
  - Normal: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$.

Cumulative Distribution Function (CDF)

- **Definition**: $F_X(x) = P(X \leq x)$, for any random variable X (discrete or continuous).
- **Properties**:
  - Non-decreasing: $F_X(x) \leq F_X(x')$ for all $x \leq x'$.

  - Right-continuous: $\lim_{y \downarrow x} F_X(y) = F_X(x)$, $y \downarrow x$ denotes that y approaches x from the right (i.e., $y \to x^+$).
  - $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
  - For discrete X: $F_X(x) = \sum_{k \leq x} p_X(k)$.
  - For continuous X: $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$.

Expectation

- **Definition**:
  - Discrete: $\mathbb{E}[X] = \sum_x x \cdot p_X(x)$.
  - Continuous: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$.
- **Linearity**: $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.
- **Law of the Unconscious Statistician (LOTUS)**: For any function $g : \mathbb{R} \to \mathbb{R}$,
  - Discrete: $\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$.
  - Continuous: $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$.

Variance and Standard Deviation

- **Variance**: $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
- **Standard Deviation**: $\sigma_X = \sqrt{\mathrm{Var}(X)}$.
- **Properties**:
  - $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.
  - $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$.

Sample Mean and Sample Variance Estimators

- **Sample Mean**: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.
  - Unbiased: $\mathbb{E}[\bar{X}_n] = \mu$.
- **Sample Variance**: $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$.
  - Unbiased: $\mathbb{E}[S_n^2] = \sigma^2$.

Confidence Interval (CI)

- **Definition**: An interval estimate for a parameter (e.g., μ) with a confidence level $(1-\alpha)$.
- **For μ (Known σ)**: CI is given by
$$\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$
where $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of $\mathcal{N}(0,1)$.
- **For μ (Unknown σ)**: CI is given by
$$\bar{X}_n \pm t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}} := \left(\bar{X}_n - t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}},\ \bar{X}_n + t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}\right),$$
where $t_{\alpha/2,n-1}$ is the quantile of the t-distribution with $n-1$ degrees of freedom.

0.3.2 Divergences and Entropies

KL-Divergence for Discrete Distributions

- **Definition**: For discrete distributions $p(x)$ and $q(x)$ with supports $\mathcal{X}_p$ and $\mathcal{X}_q$:
$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x\in\mathcal{X}_p} p(x)\log\frac{p(x)}{q(x)}. \tag{0.1}$$
  - If $\mathcal{X}_p \not\subseteq \mathcal{X}_q$, $D_{\mathrm{KL}}(p \parallel q) = +\infty$.
- **Decomposition**:
$$D_{\mathrm{KL}}(p \parallel q) = H(p,q) - H(p),$$
where:
  - **Cross Entropy**:
$$H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\log q(x). \tag{0.2}$$
  - **Entropy**:
$$H(p) = -\sum_{x\in\mathcal{X}} p(x)\log p(x). \tag{0.3}$$
- **Binary Case**:
  - Entropy: $H(p) = -\big(p_1\log p_1 + (1-p_1)\log(1-p_1)\big)$.
  - Cross Entropy: $H(p,q) = -\big(p_1\log q_1 + (1-p_1)\log(1-q_1)\big)$.
- **Properties**:
  - $D_{\mathrm{KL}}(p \parallel q) \geq 0$ with equality iff $p = q$.
  - Asymmetric: In general, $D_{\mathrm{KL}}(p \parallel q) \neq D_{\mathrm{KL}}(q \parallel p)$.

KL-Divergence for Continuous Distributions

- **Definition**: For continuous densities $p(x)$ and $q(x)$:

$$D_{\mathrm{KL}}(p \parallel q) = \int_{\mathcal{X}_p} p(x) \log \frac{p(x)}{q(x)} \, dx. \tag{0.4}$$

Jensen-Shannon Divergence

- **Definition**: Symmetric divergence for $p(x)$ and $q(x)$ with supports $\mathcal{X}_p$ and $\mathcal{X}_q$:

$$\mathrm{JSD}(p \parallel q) = \frac{1}{2} \left(D_{\mathrm{KL}}(p \parallel m) + D_{\mathrm{KL}}(q \parallel m) \right), \tag{0.5}$$

  where $m(x) = \frac{1}{2}(p(x) + q(x))$.
  - $\sqrt{\mathrm{JSD}(p \parallel q)}$ is a valid metric.

0.3.3 Computations for Multivariate Normal Distributions

Multivariate Normal Density

- **PDF**: For $x \in \mathbb{R}^m$ with mean μ and covariance Σ:

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(\det \Sigma)^{1/2} (2\pi)^{m/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)}.$$

- **Log-Density**:

$$\log \mathcal{N}(x; \mu, \Sigma) = -\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu) - \frac{m}{2} \log(2\pi) - \frac{1}{2} \log(\det \Sigma). \tag{0.6}$$

KL-Divergence for Multivariate Normals

- **General Case**: For $\mathcal{N}_{\mu_1, \Sigma_1}$ and $\mathcal{N}_{\mu_2, \Sigma_2}$:

$$D_{\mathrm{KL}}(\mathcal{N}_{\mu_1, \Sigma_1} \parallel \mathcal{N}_{\mu_2, \Sigma_2}) = \frac{1}{2} \left((\mu_1 - \mu_2)^\top \Sigma_2^{-1} (\mu_1 - \mu_2) - m + \mathrm{tr}(\Sigma_2^{-1} \Sigma_1) + \log \frac{\det \Sigma_2}{\det \Sigma_1} \right). \tag{0.7}$$

- **Special Cases**:
  - For $\Sigma_2 = \sigma_2^2 I$:

$$D_{\mathrm{KL}}(\mathcal{N}_{\mu_1, \Sigma_1} \parallel \mathcal{N}_{\mu_2, \sigma_2^2 I}) = \frac{1}{2\sigma_2^2} \|\mu_1 - \mu_2\|^2 + \frac{\mathrm{tr}(\Sigma_1)}{2\sigma_2^2} - \frac{m}{2} + \frac{m \log \sigma_2^2}{2} - \frac{\log \det \Sigma_1}{2}. \tag{0.8}$$

  - For standard normal ($\mu_2 = 0$, $\Sigma_2 = I$):

$$D_{\mathrm{KL}}(\mathcal{N}_{\mu_1, \Sigma_1} \parallel \mathcal{N}_{0, I}) = \frac{1}{2} \|\mu_1\|^2 + \frac{\mathrm{tr}(\Sigma_1)}{2} - \frac{m}{2} - \frac{\log \det \Sigma_1}{2}. \tag{0.9}$$