CS 7545: Machine Learning Theory **Fall 2021**

Lecture 1: Any Single Gaussian

Instructor: Santosh Vempala *Lecture date: 08/25/2021*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

1.1 Introduction

Given set of points $x^1, x^2, \ldots, x^m \in \mathbb{R}^d$, how to summarize data? For one-dimensional data,

- Mean $\mu = \frac{1}{m}\sum_{i=1}^m x^i$
- Variance $\sigma^2 = \frac{1}{m}\sum_{i=1}^m (x-\mu)^2$.

In higher dimensions($d > 1$), the mean is still $\mu = \frac{1}{m}\sum_{i=1}^m x^i$. The variance along any unit vector $v \in \mathbb{R}^d$, σ_v^2 is the variance of points $\langle x^1, v\rangle, \langle x^2, v\rangle, \ldots \langle x^m, v\rangle$.

$$\begin{aligned}\sigma_v^2 &= \frac{1}{m}\sum_{i=1}^m (\langle x^i, v\rangle - \langle \mu, v\rangle)^2 \\ &= \frac{1}{m}\sum_{i=1}^m \langle (x^i - \mu), v\rangle^2 \\ &= v^\top \left(\frac{1}{m}\sum_{i=1}^m (x^i-\mu)(x^i-\mu)^\top\right) v\end{aligned}$$

The matrix $\Sigma = \frac{1}{m}\sum_{i=1}^m (x^i-\mu)(x^i-\mu)^\top$ is called the *covariance matrix* of the data and $\sigma_v^2 = v^\top \Sigma v$.

$$\begin{aligned}\Sigma_{ii} &= \frac{1}{m}\sum_{l=1}^m (x_i^l - \mu_i)^2 \\ \Sigma_{ij} &= \frac{1}{m}\sum_{l=1}^m (x_i^l - \mu_i)(x_j^l - \mu_j)\end{aligned}$$

Probability Distribution

- For discrete random variable X, its probability distribution is $p(x) = \Pr(X = x)$.
- For a continuous random variable X, $p(X = x)$ is called probability density function (pdf).

Example 1.1. *Uniform density in* $[a, b]$, $p(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$.

In general, we can have a probability density function $p(x) \propto f(x)$, where $f(x) \geq 0$ and $\int f(x)dx < \infty$.

1.2 Gaussian Distribution

$\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2.

Standard Gaussian Distribution

- The pdf of 1-dimensional standard Gaussian distribution, $\mathcal{N}(0,1)$ is $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.
- The pdf of the standard (spherical) Gaussian distribution in $\mathbb{R}^d$, $\mathcal{N}(0, I_d)$ is

$$p(x) = \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{-\|x\|^2}{2}} = \prod_{i=1}^{d} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} \right).$$

  It is the product of one-dimensional Gaussian along each coordinates.

Fact. For any unit vector $v \in \mathbb{R}^d$, and $x \sim \mathcal{N}(0, I_d)$, $\langle x, v \rangle \sim \mathcal{N}(0,1)$.

Proof. Let $\{v_1, v_2, \ldots, v_d\}$ be an orthonormal basis of $\mathbb{R}^d$ with $v_1 = v$. Then $\|x\|^2 = \sum_{i=1}^{d} \langle x, v_i \rangle^2$.

$$\Pr(\langle x, v \rangle = t) = \int_{\langle x,v \rangle = t} p(x) dx = \int_{\langle x,v \rangle = t} \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{\|x\|^2}{2}} dx = \int_{\langle x,v \rangle = t} \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{\sum_{i=1}^{d} \langle x, v_i \rangle^2}{2}} dx$$

Let $y_i = \langle x, v_i \rangle$ for $i \in [d]$ and $V = (v_1 \cdots v_d)$. Then $x = Vy$ and $dx = |\det(V)| dy = dy$. So,

$$\Pr(y_1 = t) = \int_{y_1 = t} \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{\sum_{i=1}^{d} y_i^2}{2}} dy = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \cdot \int_{\mathbb{R}^{d-1}} \frac{1}{(\sqrt{2\pi})^{d-1}} e^{-\frac{\sum_{i=2}^{d} y_i^2}{2}} dy = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

□

Fact. $\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1$.

Proof. Let $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $p(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$. Then,

$$\left(\int_{\mathbb{R}} p(x) dx \right)^2 = \int_{\mathbb{R}^2} p(x,y) dx dy = \int_{\mathbb{R}^2} \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} dx dy.$$

For the change of variables for $x = r\cos\theta, y = r\sin\theta$, we get $dxdy = 2\pi r dr$, and

$$\left(\int_{\mathbb{R}} p(x) dx \right)^2 = \int_{\mathbb{R}} \frac{2\pi r}{2\pi} e^{-\frac{r^2}{2}} dr = 1.$$

□

General Gaussian Distribution

- The pdf of one-dimesional Gaussian $\mathcal{N}(\mu, \sigma^2)$ is $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- The pdf of Gaussian $\mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^d$ is

$$p(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det \Sigma}} e^{-\frac{(x-\mu)^\top \Sigma^{-1} (x-\mu)}{2}}$$

Let $Y = AX + b$, where $X \sim \mathcal{N}(0, I_d)$, then $Y \sim \mathcal{N}(b, AA^\top)$.

$$\mathbb{E}(Y) = \mathbb{E}(AX + b) = b \text{ and } \mathrm{Var}(y) = \mathbb{E}((AX)(AX)^\top) = A\mathbb{E}(XX^\top)A^\top = AA^\top.$$

Fact. If $X \sim \mathcal{N}(\mu, \Sigma)$ and $Y = \Sigma^{-1/2}(X - \mu)$, then $Y \sim \mathcal{N}(0, I_d)$.

1.2.1 Best fit Gaussian

Problem 1. *Given random i.i.d. samples $x^1, x^2, \ldots, x^m$, estimate the parameters (μ, Σ) of a Gaussian that maximizes the likelihood of generating the observed samples.*

A natural choice for (μ, Σ) would be the empirical mean, $\tilde{\mu} = \frac{1}{m}\sum_{i=1}^m x_i$ and the empirical covariance matrix, $\tilde{\Sigma} = \frac{1}{m}\sum_{i=1}^m (x^i - \tilde{\mu})(x^i - \tilde{\mu})^\top$.

Theorem 1.2. *Given random i.i.d. samples $x^1, x^2, \ldots, x^m$, the emprical mean and covariance, $(\tilde{\mu}, \tilde{\Sigma})$, are the maximum likelihood estimators.*

Proof. Consider the case when $\Sigma = \sigma^2 I_d$. The probability of observing $x^1, x^2, \ldots, x^m$ is

$$p(x^1, x^2, \ldots, x^m) = \prod_{i=1}^m \left(\frac{1}{(\sqrt{2\pi}\sigma)^d} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}} \right).$$

Taking log, we get the log-likelihood function,

$$L(\mu, \sigma) = \log p(x^1, x^2, \ldots, x^m; \mu, \sigma) = -md \log \sqrt{2\pi}\sigma - \sum_{i=1}^m \frac{\left\|x^i - \mu\right\|^2}{2\sigma^2}.$$

Taking partial derivatives of $L(\mu, \sigma)$, we get

$$\frac{\partial L(\mu,\sigma)}{\partial \mu_j} = -\frac{1}{\sigma^2}\sum_{i=1}^m (x_j^i - \mu_j), \quad \frac{\partial L(\mu,\sigma)}{\partial \sigma} = -\frac{md}{\sigma} + \frac{\sum_{i=1}^m \left\|x^i - \mu\right\|^2}{\sigma^3}.$$

$$\frac{\partial L(\mu,\sigma)}{\partial \mu_j} = 0 \Rightarrow \mu_j = \frac{1}{m}\sum_{i=1}^m x_j^i, \text{ and}$$
$$\frac{\partial L(\mu,\sigma)}{\partial \sigma} = 0 \Rightarrow \sigma^2 = \frac{1}{md}\sum_{i=1}^m \|x - \mu\|^2$$

Therefore, $\mu = \frac{1}{m}\sum_{i=1}^m x^i$ and $\sigma^2 = \frac{1}{md}\sum_{i=1}^m \|x-\mu\|^2$ maximize $L(\mu, \sigma)$. □

Theorem 1.3 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables drawn from some distribution with bounded variance,* $\mathrm{Var}(X_i) \le \infty$*, then $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then,*

$$Y_n \xrightarrow{\textit{converges in distribution}} \mathcal{N}\left(\mathbb{E}(X_1), \frac{1}{n}\mathrm{Var}(X_1)\right).$$

Theorem 1.4 (Berry-Esseen theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables. Let $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $Z_n \sim \mathcal{N}(\mathbb{E}(Y_n), \mathrm{Var}(Y_n))$. Then, for all $t \in \mathbb{R}$,*

$$|\Pr(Y_n \le t) - \Pr(Z_n \le t)| \le \frac{c \cdot \sum_{i=1}^n \mathbb{E}(|X_i|^3)}{\mathrm{Var}(Y_n)^{3/2}},$$

where c is an absolute constant in $[0, 1]$.

Example 1.5. *Let* $X_i = \begin{cases} -1 & w.p.\ 1/2 \\ 1 & w.p.\ 1/2 \end{cases}$ *and $Y_n = \sum_{i=1}^n X_i$.*

Then $\mathbb{E}(X_i) = 0$, $\mathbb{E}(X_i^2) = \mathbb{E}(|X_i^3|) = 1$, $\mathbb{E}(Y_n) = 0$, and $\mathrm{Var}(Y_n) = n$. So, for $Z_n \sim \mathcal{N}(0, n)$ and for any $t \in \mathbb{R}$,

$$|\Pr(Y_n \le t) - \Pr(Z_n \le t)| \le \frac{cn}{n^{3/2}} = cn^{-1/2}.$$