COMP 7070: Advanced Topics in Artificial Intelligence and Machine Learning

Lecture 1: Preliminaries

Instructor: Yifan Chen *Scribes: Wenbo Shang, Rui Cao* *Date: January 9, 2024*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Acknowledgement: We extend our sincere gratitude to Zhanke Zhou for his meticulous proofreading of this manuscript.

1.1 Gaussian Distribution

The Gaussian distribution, also known as the normal distribution, is a continuous probability distribution that is symmetric around its mean. It is characterized by its mean (μ) and standard deviation (σ). In this section, we study the matrix form of the distribution with the random variable $X \in \mathbb{R}^{n\times 1}$ below:

$$p(x_i; \mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^n} \cdot \frac{1}{|\Sigma|^{\frac{1}{2}}} \cdot exp(-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)) \tag{1.1}$$

1.1.1 T-test

Assume $X_n \sim N(\mu \cdot \mathbf{1}, \sigma^2 \mathbf{I})$, where $\mathbf{1}$ is a vector of ones in the form of $(1, 1, ..., 1)^\top_{n\times 1}$, $\mathbf{I}$ is an identity matrix, and σ^2 is unknown. Let X_n^b and X_n^y be two random independent variables that both obey the distribution in Equation 1.1. Under the condition, we make the hypothesis $\mu_y > \mu_b$. As a result, we obtain a new random variable $X_n^y - X_n^b$ satisfying $(X_n^y - X_n^b) \sim N(0, (\sigma_y^2 + \sigma_b^2)\mathbf{I})$.

1.1.2 T-distribution

We present the basic form of T distribution here: $T = \frac{z}{\sqrt{\frac{s}{d}}}$, where random variables z and s satisfy: ①$z \sim N(0, 1)$. ②$s \sim \chi^2(d)$, where d is the degree of freedom of the distribution. ③ z and s are independent of each other. More generally, we have the representation below:

$$T = \frac{(\bar{X} - \mu)}{\hat{\sigma}/\sqrt{n}} \sim T(n-1), where\ \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2. \tag{1.2}$$

Proof. We can first rewrite the Equation 1.2 like the basic form of T distribution:

$$T = \frac{(\bar{X}-\mu)/\sqrt{\sigma^2/n}}{\hat{\sigma}/\sqrt{n}/\sqrt{\sigma^2/n}} \triangleq \frac{z}{\sqrt{\frac{s}{n-1}}} \tag{1.3}$$

We can easily find that condition ① has been proved since the numerator in Equation 1.3 satisfies the normal distribution. Next, we want to prove the correctness of condition ②.

$$\begin{aligned} s &= \frac{n-1}{\sigma^2}\hat{\sigma}^2 = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n}(x_i - \bar{X})^2 \\ &= \frac{1}{\sigma^2} \cdot (X - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^\top X)^\top (X - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^\top X) \\ &= \frac{1}{\sigma^2} \cdot [(\mathbf{I} - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^\top)X)]^\top [(\mathbf{I} - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^\top)X] \end{aligned} \tag{1.4}$$

Let $P := \mathbf{I} - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^\top$, thus P is a projection matrix which can be rewritten as below:

$$P = \bar{U} \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \bar{U}^\top = UU^\top, where\ U \in \mathbb{R}^{n\times(n-1)}. \tag{1.5}$$

Based on Equation 1.4 and 1.5, we have:

$$s = \frac{1}{\sigma^2} \cdot X^\top P^\top PX = \frac{1}{\sigma^2} X^\top U(U^\top X) \tag{1.6}$$

Let $Y_{n-1} := U^\top X \sim N(\mu \cdot U^\top \cdot \mathbf{1}, \sigma^2 \cdot U^\top U)$, where $U^\top U = \mathbf{I}_{n-1}$. On the other hand, we can infer that $P \cdot \mathbf{1} = 0$ according to the definition of P and the characteristics of the projection matrix, thus $U^\top \mathbf{1} = 0$. Finally, we have:

$$s = \frac{1}{\sigma^2} X^\top U(U^\top X) = \frac{1}{\sigma^2} Y_{n-1}^\top Y_{n-1} = \frac{1}{\sigma^2} \sum_{i=1}^{n-1} Y_i^2 \sim \chi^2(n-1) \tag{1.7}$$

The condition ② has been proved above. In terms of the independence between z and s, we just need to calculate the covariance between $\bar{X}$ and Y based on their definition.

$$\begin{aligned} Cov(Y, \bar{X}) &= EY \cdot \bar{X} - EY \cdot E\bar{X} \xlongequal{EY=0} EU^\top X \cdot \frac{1}{n} \cdot X^\top \cdot \mathbf{1} \\ &= \frac{1}{n} U^\top (EXX^\top) \cdot \mathbf{1} \\ &= \frac{1}{n} U^\top [(\mu \cdot \mathbf{1})(\mu \cdot \mathbf{1})^\top + \sigma^2 \mathbf{I}] \cdot \mathbf{1} \\ &= \frac{1}{n} U^\top \mu^2 \cdot \mathbf{1} \cdot \mathbf{1}^\top \cdot \mathbf{1} + \frac{1}{n} U^\top \sigma^2 \cdot \mathbf{I} \cdot \mathbf{1} = \mathbf{0}_{n-1} \end{aligned} \tag{1.8}$$

The Equation 1.8 shows that condition ③ is correct thus Equation 1.2 is true. The T-distribution is commonly used in hypothesis testing and in the construction of t-tests. It allows for inference about population means when the population standard deviation is unknown here. □

1.1.3 Maximum Likelihood Estimator

$\bar{X}$ as an estimator for μ. Formula of Maximum Likelihood Estimator(MLE):

$$\bar{X} = \arg\max_{\mu} \prod_{i=0}^{n} p(x_i; \mu, \Sigma) \tag{1.9}$$

Next step, we want to transform $\prod$ to $\sum$. It's a normal trick to transform complex problem to a better solved problem. For example, we use log to transform e^{-x^2} to $-x^2$ from a non-concave function to a concave function. Similarly, MLE can also be transformed as below using log trick.

$$\sum \ln p(x; \mu, \Sigma) \propto \sum_{i=0}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \tag{1.10}$$

Next, we use matrix trace to simplify the above problem.

$$\begin{aligned} &\min \sum_{i=0}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \\ &\iff \min Tr(X_{n\times d} - \mathbf{1} \cdot \mu^T)^T \Sigma^{-1} (X_{n\times d} - \mathbf{1} \cdot \mu^T) \\ &:= f(\mu) \end{aligned} \tag{1.11}$$

After defining f, we derive f like:

$$df = Tr((\frac{\partial f}{\partial \mu})^T du) \tag{1.12}$$

To calculate the derivative of f, let's recall how to derive the matrix trace firstly. For matrix $\boldsymbol{ABC}$, the trace of product of matrix $\boldsymbol{ABC}$ denotes as $Tr(\boldsymbol{ABC})$.The derivative of the trace is:

$$df = Tr[\boldsymbol{dA} \cdot \boldsymbol{BC} + \boldsymbol{A} \cdot \boldsymbol{dB} \cdot \boldsymbol{C} + \boldsymbol{AB} \cdot \boldsymbol{dC}] \tag{1.13}$$

In addition, based on properties of matrix traces, we know that:

$$Tr(\boldsymbol{AB}) = Tr(\boldsymbol{BA}) \tag{1.14}$$

Based on 1.13 and 1.14, the derivative of the MLE can be deduced as below:

$$df = Tr[d(X - \mathbf{1} \cdot \mu^T)^T \Sigma^{-1}(X - \mathbf{1} \cdot \mu^T) + 0 + (X - \mathbf{1} \cdot \mu^T)^T \Sigma^{-1} d(X - \mathbf{1} \cdot \mu^T)] \tag{1.15}$$

Let $\boldsymbol{A}^T = \Sigma^{-1}(X - \mathbf{1} \cdot \mu^T)$:

$$\begin{aligned} df &= Tr[d(X - \mathbf{1} \cdot \mu^T)^T \cdot \boldsymbol{A}^T + 0 + \boldsymbol{A} \cdot d(X - \mathbf{1} \cdot \mu^T)] \\ &= Tr[-\mathbf{1} \cdot d\mu^T \cdot \boldsymbol{A}^T + \boldsymbol{A} \cdot (-d\mu) \cdot -\mathbf{1}^T] \\ &= Tr[-2 \cdot -\mathbf{1} \cdot \boldsymbol{A} \cdot d\mu] \end{aligned} \tag{1.16}$$

Therefore, we derive the first derivative of f. Since f is a convex function, when the value of f is the smallest, it is the global minimum point. At this point, the first-order derivative is equal to 0, so we can deduce:

$$\begin{aligned} \frac{\partial f}{\partial \mu} &= -2 \cdot \boldsymbol{A} \cdot \mathbf{1} = 0 \\ &\Rightarrow \Sigma^{-1}(X - \mathbf{1} \cdot \mu^T) \cdot \mathbf{1} = 0 \\ &\Rightarrow \mu = \frac{1}{n} \cdot X^T \cdot \mathbf{1} \end{aligned} \tag{1.17}$$

Finally, we get the value of μ.

1.2 Linear Regression

1.2.1 linear model

Assume we have a linear model which can be described as the equation below:

$$Y = \begin{pmatrix} 1 & X^{n \times d} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1^{d \times 1} \end{pmatrix} + \mathbf{e}. \tag{1.18}$$

In the equation above, $X^{n \times d}$ is independent variable and β_0, β_1 are coefficient term. The constant term $\mathbf{e}$ of Equation 1.18 representing the error must satisfy Gaussian-Markov condition:①E $\mathbf{e}$=0.②Var $\mathbf{e} = \sigma^2 \mathbf{I}^{n \times n}$.

1.2.2 Square Loss

Square loss is:

$$\frac{1}{2n}(Y - \bar{X}\bar{\beta})^T(Y - \bar{X}\bar{\beta}), where\ \bar{X} = (1, X), \bar{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}. \tag{1.19}$$

If we want to get the minimum square loss, we also need to find the global minimum point. We all know that the square loss function is a convex function. Therefore, the point where the function's first derivative equals 0 is the global minimum point. Assume the square loss function's first derivative equals 0, we can get the value of $\hat{\bar{\beta}}$:

$$\begin{aligned} &\frac{1}{n}\bar{X}^T(Y - \bar{X}\hat{\bar{\beta}}) = 0 \\ \Rightarrow\ &\hat{\bar{\beta}} = (\bar{X^T}\bar{X})^{-1}\bar{X}^T Y \end{aligned} \tag{1.20}$$

Lemma 1.1. *$\hat{\bar{\beta}}$ is the MLE for Gaussian variable e.*

Proof. Firstly, we prove the expectation of $\hat{\bar{\beta}}$. For $Y = \bar{X}\bar{\beta} + e$, we can deduce that:

$$\begin{aligned} E\hat{\bar{\beta}} &= E(\bar{X}^T\bar{X})^{-1}\bar{X}^T(\bar{X}\bar{\beta} + e) \\ &= (\bar{X}^T\bar{X})^{-1}(\bar{X}^T\bar{X})\bar{\beta} + (\bar{X}^T\bar{X})^{-1}(\bar{X}^T\bar{X})Ee \end{aligned} \tag{1.21}$$

Based on Gaussian-Markov condition, we know that $Ee = 0$, thus, $(\bar{X}^T\bar{X})^{-1}(\bar{X}^T\bar{X})Ee = 0$. We can deduce that:

$$\begin{aligned} E\hat{\bar{\beta}} &= (\bar{X}^T\bar{X})^{-1}(\bar{X}^T\bar{X})\bar{\beta} \\ &= \bar{\beta} \end{aligned} \tag{1.22}$$

∎

Proof. Secondly, we prove the variance of $\hat{\bar{\beta}}$. For $Var(AY) = A \cdot Var(Y)A^T$, we can deduce that:

$$\begin{aligned} Var(\hat{\bar{\beta}}) &= (\bar{X}^\top\bar{X})^{-1}\bar{X}^\top \cdot Var(Y) \cdot \bar{X}(\bar{X}^\top\bar{X})^{-1} \\ &= \sigma^2(\bar{X}^\top\bar{X})^{-1}\bar{X}^\top\bar{X}(\bar{X}^\top\bar{X})^{-1} \\ &= \sigma^2(\bar{X}^\top\bar{X})^{-1} \end{aligned} \tag{1.23}$$

∎

We prove $E\hat{\bar{\beta}} = \bar{\beta}$ and $Var(\hat{\bar{\beta}}) = \sigma^2(\bar{X}^\top\bar{X})^{-1}$, thus, $\hat{\bar{\beta}}$ is the MLE for Gaussian variable e.

1.2.3 Population risk

In this section, we evaluate the population risk of our linear model. We remark during the training process, we are minimizing the *empirical risk* on fixed design $\bar{X}$.

Specifically, we assume a new random sample $\boldsymbol{x}$ and the corresponding label $\boldsymbol{y} = \boldsymbol{x}^\top\beta + \boldsymbol{\varepsilon}$; the risk is expressed as below:

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y},\hat{\bar{\beta}}}(\boldsymbol{y} - \boldsymbol{x}^\top\hat{\bar{\beta}})^2 = \mathbb{E}_{\hat{\bar{\beta}}}[\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{y} - \boldsymbol{x}^\top\hat{\bar{\beta}})^2 \mid \hat{\bar{\beta}}] \tag{1.24}$$

Considering the noise $\boldsymbol{\varepsilon}$ within $\boldsymbol{y}$ is independent from all the other variables, and we cannot predict this part, the preceding population risk indeed depends on $\mathbb{E}_{\boldsymbol{x},\hat{\beta}}\boldsymbol{x}^\top(\beta - \hat{\beta})(\beta - \hat{\beta})^\top\boldsymbol{x}$ (we abuse $\hat{\beta}$ as $\hat{\bar{\beta}}$ for simplicity from then on). We are equivalently minimizing

$$\begin{aligned} \min \mathbb{E}_{\boldsymbol{x},\hat{\beta}}\boldsymbol{x}^\top(\beta - \hat{\beta})(\beta - \hat{\beta})^\top\boldsymbol{x} &= \mathbb{E}_{\hat{\beta}}[\mathbb{E}_{\boldsymbol{x}}\mathrm{Tr}[\boldsymbol{x}\boldsymbol{x}^\top(\beta - \hat{\beta})(\beta - \hat{\beta})^\top] \mid \hat{\beta}] \\ &= \mathbb{E}_{\hat{\beta}}[\mathrm{Tr}[\mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top(\beta - \hat{\beta})(\beta - \hat{\beta})^\top] \mid \hat{\beta}] \\ &= \mathrm{Tr}[\mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top \cdot \mathbb{E}(\beta - \hat{\beta})(\beta - \hat{\beta})^\top] \\ &= \sigma^2\mathrm{Tr}[\mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top(\bar{X}^\top\bar{X})^{-1}] \end{aligned} \tag{1.25}$$

In Equation 1.25, we recall $\bar{X}$ is the fixed sample value in the training process. In other words, we are able to manipulate $\bar{X}$ to minimize the population risk, as other quantities are fixed ($\mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top$ is fixed when $\boldsymbol{x}$ is given). Assuming $\mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top = \boldsymbol{I}$, in this case the preceding display reads $\sigma^2\mathrm{Tr}[(\bar{X}^\top\bar{X})^{-1}] = \mathrm{Tr}[\mathrm{var}(\hat{\beta})]$, indicating the connection between population risk and parameter variance in the case of linear regression.