CS229: Machine Learning

Tianyu Du

June 29, 2019

Contents

1 Lecture Notes Jun. 24 2019

1.1 Review of Linear Algebra

Remark 1.1. In this course, vectors are treated as *column matrices*.

Definition 1.1. Given $A \in M_{n\times n}(\mathbb{R})$, the trace of A is defined as

$$\operatorname{tr}(A) := \sum_{i=1}^{n} A_{i,i} \tag{1.1}$$

Definition 1.2. Given $x, y \in \mathbb{R}^n$, the **inner product** is defined as

$$\langle x, y\rangle := x^T y = \sum_{i=1}^{n} x_i \; y_i \tag{1.2}$$

Definition 1.3. Given $x \in \mathbb{R}^b, y \in \mathbb{R}^p$, the **outer product** is defined as

$$x \otimes y := xy^T = A \in M_{b\times p}(\mathbb{R}) \tag{1.3}$$

in which

$$A_{i,j} := x_i \; y_j \tag{1.4}$$

the constructed matrix A is a **rank 1 matrix**.

Remark 1.2. Given two rank 1 matrices A_1 and A_2, then $A_1 + A_2$ is a rank 2 matrix.

Remark 1.3. Note that the outer product operation is not commutative.

Definition 1.4. Let $v, b \in \mathbb{R}^n$, the **projection matrix** of v is defined as $\frac{vv^T}{v^Tv} \equiv \frac{v\otimes v}{\langle v,v\rangle}$. Then $\frac{v\otimes v}{\langle v,v\rangle}b$ is the projection of b on v.

$$\begin{aligned} \frac{v \otimes v}{\langle v, v\rangle} b &= \left[\frac{v}{\langle v, v\rangle}\right]\left[\frac{v}{\langle v, v\rangle}\right]^T b && (1.5)\\ &= \tilde{v} \; \underbrace{\tilde{v}^T b}_{\text{magnitude}} && (1.6) \end{aligned}$$

Proposition 1.1. Let $A \in M_{m\times n}(\mathbb{R})$, the projection of vector $b \in \mathbb{R}^m$ onto the *column space* of A is given by the generalized projection matrix

$$A(A^TA)^{-1}A^Tb \tag{1.7}$$

2 Lecture Notes Jun. 28 2019

Example 2.1 (Maximum Likelihood Estimation for Multivariate Gaussian Distribution)**.**

Lemma 2.1. Let $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$, then

$$\nabla_A x^T A x = xx^T \tag{2.1}$$

Proof.

$$x^T A x = \begin{pmatrix} \sum_{i=1}^n x_i A_{i,1} \\ \vdots \\ \sum_{i=1}^n x_i A_{i,n} \end{pmatrix} x = \sum_{j=1}^n \sum_{i=1}^n x_i A_{i,j} x_j \tag{2.2}$$

$$\implies \nabla_A x^T A x_{i,j} = \frac{\partial \sum_{j=1}^n \sum_{i=1}^n x_i A_{i,j} x_j}{\partial A_{i,j}} = x_i x_j \tag{2.3}$$

$$\implies \nabla_A x^T A x = x x^T \tag{2.4}$$

■

Lemma 2.2. Let $x \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n\times n}$, then

$$\nabla_x x^T A x = 2x^T A \tag{2.5}$$

Lemma 2.3. Let $A \in \mathbb{R}^{n\times n}$ such that A is non-singular, then

$$\nabla_A \ln(|A|) = A^{-1} \tag{2.6}$$

Derive the MLE for Gaussian. Let $\left(x^{(i)}\right)_{i=1}^n$ denote the set of training instances. Assuming they are independently and identically distributed (*i.i.d.*) following $\mathcal{N}(\mu, \Sigma)$, the joint likelihood can be written as

$$\mathcal{L}(\mu, \Sigma; x^{(i)}) = \prod_{i\in[n]} \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)\right) \tag{2.7}$$

Then the MLE becomes the maximizer of the log-likelihood

$$(\hat{\mu}, \hat{\Sigma}) := \operatorname*{argmax}_{\mu,\Sigma} \ell(\mu, \Sigma; x^{(i)}) \tag{2.8}$$

$$= \operatorname*{argmax}_{\mu,\Sigma} \sum_{i\in[n]} \left\{ \ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu) \right\} \tag{2.9}$$

Then the first order condition for $\hat{\mu}$ is

$$\nabla_\mu \ell(\mu, \Sigma, x^{(i)})|_{\mu=\hat{\mu}} = 0 \tag{2.10}$$

$$\implies \sum_{i\in[n]} (x^{(i)} - \hat{\mu})^T \Sigma^{-1} = 0 \tag{2.11}$$

$$\implies \Sigma^{-1} n \hat{\mu} = \Sigma^{-1} \sum_{i\in[n]} x^{(i)} \tag{2.12}$$

$$\implies \hat{\mu} = \frac{1}{n} \sum_{i\in[n]} x^{(i)} \tag{2.13}$$

For $\hat{\Sigma}$, define $S := \Sigma^{-1}$, note that $\nabla_S \ell = 0 \iff \nabla_{\Sigma^{-1}} \ell = 0$

$$\begin{aligned}
&\nabla_S = 0 && (2.14)\\
\implies &\nabla_S \sum_{i\in[n]} \left\{ \frac{1}{2} \ln(|S|) - \frac{1}{2}(x^{(i)} - \mu)^T S (x^{(i)} - \mu) \right\} = 0 && (2.15)\\
\implies &\sum_{i\in[n]} \left\{ S^{-1} - (x^{(i)} - \mu)(x^{(i)} - \mu)^T \right\} = 0 && (2.16)\\
\implies &S^{-1} = \sum_{i\in[n]} (x^{(i)} - \mu)(x^{(i)} - \mu)^T && (2.17)\\
\implies &\hat{\Sigma} = \frac{1}{n} \sum_{i\in[n]} (x^{(i)} - \mu)(x^{(i)} - \mu)^T \approx \mathbb{E}[(x^{(i)} - \mathbb{E}[x^{(i)}])^2] \equiv \mathbb{V}[x^{(i)}] && (2.18)
\end{aligned}$$

■