CS229: Machine Learning

Tianyu Du

July 1, 2019

Contents

Chapter 1

Preliminary

1.1 Lecture Notes Jun. 24 2019

1.1.1 Review of Linear Algebra

Remark 1.1.1. In this course, vectors are treated as *column matrices*.

Definition 1.1.1. Given $A \in M_{n\times n}(\mathbb{R})$, the trace of A is defined as

$$\operatorname{tr}(A) := \sum_{i=1}^{n} A_{i,i} \tag{1.1.1}$$

Definition 1.1.2. Given $x, y \in \mathbb{R}^n$, the **inner product** is defined as

$$\langle x, y\rangle := x^T y = \sum_{i=1}^{n} x_i\ y_i \tag{1.1.2}$$

Definition 1.1.3. Given $x \in \mathbb{R}^b, y \in \mathbb{R}^p$, the **outer product** is defined as

$$x \otimes y := xy^T = A \in M_{b\times p}(\mathbb{R}) \tag{1.1.3}$$

in which

$$A_{i,j} := x_i\ y_j \tag{1.1.4}$$

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

Remark 1.1.2. Given two rank 1 matrices A_1 and A_2, then $A_1 + A_2$ is a rank 2 matrix.

Remark 1.1.3. Note that the outer product operation is not commutative.

Definition 1.1.4. Let $v, b \in \mathbb{R}^n$, the **projection matrix** of v is defined as $\frac{vv^T}{v^T v} \equiv \frac{v\otimes v}{\langle v,v\rangle}$. Then $\frac{v\otimes v}{\langle v,v\rangle} b$ is the projection of b on v.

$$\frac{v\otimes v}{\langle v,v\rangle} b = \left[\frac{v}{\langle v,v\rangle}\right]\left[\frac{v}{\langle v,v\rangle}\right]^T b \tag{1.1.5}$$

$$= \tilde{v}\ \underbrace{\tilde{v}^T b}_{\text{magnitude}} \tag{1.1.6}$$

Proposition 1.1.1. 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

$$A(A^TA)^{-1}A^Tb \tag{1.1.7}$$

1.2 Lecture Notes Jun. 28 2019

Example 1.2.1 (Maximum Likelihood Estimation for Multivariate Gaussian Distribution)**.**

Lemma 1.2.1. Let $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$, then

$$\nabla_A x^TAx = xx^T \tag{1.2.1}$$

Proof.

$$x^TAx = \begin{pmatrix} \sum_{i=1}^n x_iA_{i,1} \\ \vdots \\ \sum_{i=1}^n x_iA_{i,n} \end{pmatrix} x = \sum_{j=1}^n\sum_{i=1}^n x_iA_{i,j}x_j \tag{1.2.2}$$

$$\implies \nabla_A x^TAx_{i,j} = \frac{\partial \sum_{j=1}^n \sum_{i=1}^n x_iA_{i,j}x_j}{\partial A_{i,j}} = x_ix_j \tag{1.2.3}$$

$$\implies \nabla_A x^TAx = xx^T \tag{1.2.4}$$

■

Lemma 1.2.2. Let $x \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n\times n}$, then

$$\nabla_x x^TAx = 2x^TA \tag{1.2.5}$$

Lemma 1.2.3. Let $A \in \mathbb{R}^{n\times n}$ such that A is non-singular, then

$$\nabla_A \ln(|A|) = A^{-1} \tag{1.2.6}$$

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

$$\mathcal{L}(\mu, \Sigma; x^{(i)}) = \prod_{i\in[n]} \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu)\right) \tag{1.2.7}$$

Then the MLE becomes the maximizer of the log-likelihood

$$(\hat{\mu}, \hat{\Sigma}) := \operatorname*{argmax}_{\mu,\Sigma} \ell(\mu, \Sigma; x^{(i)}) \tag{1.2.8}$$

$$= \operatorname*{argmax}_{\mu,\Sigma} \sum_{i\in[n]} \left\{ \ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu) \right\} \tag{1.2.9}$$

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

$$\nabla_\mu \ell(\mu, \Sigma, x^{(i)})|_{\mu=\hat{\mu}} = 0 \tag{1.2.10}$$

$$\implies \sum_{i\in[n]} (x^{(i)} - \hat{\mu})^T \Sigma^{-1} = 0 \tag{1.2.11}$$

$$\implies \Sigma^{-1} n\hat{\mu} = \Sigma^{-1} \sum_{i\in[n]} x^{(i)} \tag{1.2.12}$$

$$\implies \hat{\mu} = \frac{1}{n} \sum_{i\in[n]} x^{(i)} \tag{1.2.13}$$

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

$$\nabla_S = 0 \tag{1.2.14}$$

$$\implies \nabla_S \sum_{i\in[n]} \left\{ \frac{1}{2} \ln(|S|) - \frac{1}{2}(x^{(i)} - \mu)^T S (x^{(i)} - \mu) \right\} = 0 \tag{1.2.15}$$

$$\implies \sum_{i\in[n]} \left\{ S^{-1} - (x^{(i)} - \mu)(x^{(i)} - \mu)^T \right\} = 0 \tag{1.2.16}$$

$$\implies S^{-1} = \sum_{i\in[n]} (x^{(i)} - \mu)(x^{(i)} - \mu)^T \tag{1.2.17}$$

$$\implies \hat{\Sigma} = \frac{1}{n} \sum_{i\in[n]} (x^{(i)} - \mu)(x^{(i)} - \mu)^T \approx \mathbb{E}[(x^{(i)} - \mathbb{E}[x^{(i)}])^2] \equiv \mathbb{V}[x^{(i)}] \tag{1.2.18}$$

■

Chapter 2

Supervised Learning

2.1 Lecture Notes Jul. 1 2019

Supervised Learning Problem Learn a mapping from input space to output space $f : X \to Y$. Given a data set $(x_i, y_i)_i$, in regression problem, the algorithm is trying to learn a *hypothesis* $h(x) \approx f$.

Terminologies

(i) n: number of (x, y) examples;

(ii) d: $x \in \mathbb{R}^d$;

(iii) $x^{(i)}$: input;

(iv) $y^{(i)}$: label or output;

(v) $(x^{(i)}, y^{(i)})$: i^{th} example.

2.1.1 Linear Regression

Model

$$h_\theta(x) = \theta_0 + \sum_{i=1}^{d} \theta_i x_i,\ \theta \in \mathbb{R}^{d+1} \tag{2.1.1}$$

$$= \langle \theta, x \rangle,\ x := (1, x_1, x_2, \ldots, x_d) \in \mathbb{R}^{d+1} \tag{2.1.2}$$

Cost/Loss Function

$$J(\theta) := \frac{1}{2} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \tag{2.1.3}$$

$$\hat{\theta} := \operatorname*{argmin}_{\theta \in \mathbb{R}^{d+1}} J(\theta) \subset \mathbb{R}^{d+1} =: \Theta \tag{2.1.4}$$

$$= \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{2} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \tag{2.1.5}$$

(Full) Gradient Descent

$$\theta^{(t+1)} := \theta^{(t)} - \alpha \nabla_\theta J(\theta^{(t)}) \tag{2.1.6}$$

where α denotes the **learning rate**. With appropriate learning rate

$$J(\theta^{(t+1)}) < J(\theta^{(t)}) \tag{2.1.7}$$

And the updating step is repeated till it converges. Theoretically, the *convergence* means the convergence of series $(\theta^{(t)})_i$.

Practical ways to check convergence

(i) $||\theta^{(t+1)} - \theta^{(t)}|| < \varepsilon$;

(ii) $\left|J(\theta^{(t+1)}) - J(\theta^{(t)})\right| < \varepsilon$;

(iii) $||\nabla_\theta J(\theta^{(t)})|| < \varepsilon$.

Gradient Descent on Linear Regression

1. Initialize $\theta^{(0)}$;
2. Repeat till convergence:

$$\theta^{(t+1)} := \theta^{(t)} - \alpha \nabla_\theta J(\theta^{(t)}) \tag{2.1.8}$$

$$= \theta^{(t)} - \alpha \nabla_\theta \frac{1}{2} \sum_{i=1}^{n} \left(\langle \theta^{(t)}, x^{(i)} \rangle - y^{(i)} \right)^2 \tag{2.1.9}$$

$$= \theta^{(t)} - \alpha \sum_{i=1}^{n} \left(\langle \theta^{(t)}, x^{(i)} \rangle - y^{(i)} \right) x^{(i)} \tag{2.1.10}$$

Stochastic Gradient Descent (SGD) Gradient descent based on one randomly selected training sample k, and use a proxy loss function:

$$\tilde{J}(\theta) := \frac{1}{2} \left(\langle \theta^{(t)}, x^{(k)} \rangle - y^{(k)} \right)^2 \tag{2.1.11}$$

and update

$$\theta^{(t+1)} := \theta^{(t)} - \alpha \left(\langle \theta^{(t)}, x^{(k)} \rangle - y^{(k)} \right) x^{(k)} \tag{2.1.12}$$

where the $\theta^{(t)}$ will finally in a region near the global minimum, the size of region is characterized by the learning rate α.

Remark 2.1.1. SGD takes more iterations to converge than full gradient descent, but the computational cost of each step is much less than full GD.

Analytical Solution to OLS

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(\theta^T x^{(i)} - y^{(i)} \right) \tag{2.1.13}$$

A **design matrix**, $\mathbf{X}$, is a $n \times d$ matrix, in which each row is an input vector. $y \in \mathbb{R}^{n \times 1}$ denotes the collection of labels. Then the loss function can be expressed as

$$J(\theta) = \frac{1}{2}||\mathbf{X}\theta - y||^2 = \frac{1}{2}[\mathbf{X}\theta - y]^T[\mathbf{X}\theta - y] \tag{2.1.14}$$

The OLS estimator can be obtained by solving

$$\nabla_\theta J(\theta) = 0 \tag{2.1.15}$$
$$\iff \nabla_\theta \frac{1}{2}||\mathbf{X}\theta - y|| = 0 \tag{2.1.16}$$
$$\iff \frac{1}{2}2[\mathbf{X}\theta - y]^T\mathbf{X} = 0 \tag{2.1.17}$$
$$\iff \theta^T\mathbf{X}^T\mathbf{X} - y^T\mathbf{X} = 0 \tag{2.1.18}$$
$$\iff \mathbf{X}^T\mathbf{X}\theta - \mathbf{X}^T y = 0 \quad \textbf{(normal equation)} \tag{2.1.19}$$
$$\implies \hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \tag{2.1.20}$$

Probabilistic Interpretation Assuming the underlying data generating process follows

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)} \tag{2.1.21}$$
$$\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2) \tag{2.1.22}$$
$$\implies y^{(i)}|x^{(i)} \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2) \tag{2.1.23}$$
$$\implies P(y^{(i)}|x^{(i)};\theta) = \phi(y^{(i)}; x^{(i)}, \theta) \tag{2.1.24}$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \tag{2.1.25}$$

Maximum Likelihood Estimation Assuming $\varepsilon^{(i)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, then the log-likelihood can be expressed as

$$\ell(\theta; x^{(i)}, y^{(i)}) = \prod_{i=1}^{n} P(y^{(i)}|x^{(i)};\theta) \tag{2.1.26}$$
$$= \sum_{i=1}^{n} \ln\left[P(y^{(i)}|x^{(i)};\theta)\right] \tag{2.1.27}$$
$$= \sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)\right] \tag{2.1.28}$$
$$= \sum_{i=1}^{n} K - \frac{1}{2\sigma^2}(y^{(i)} - \theta^T x^{(i)})^2 \tag{2.1.29}$$
$$= \tilde{K} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \theta^T x^{(i)})^2 \tag{2.1.30}$$

Conclusion The ordinary least square estimator coincides with the maximum likelihood estimator. If we assume the noise in the model follows Gaussian distribution, then maximizing the likelihood and minimizing loss function are equivalent:

$$\underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell(\theta; x^{(i)}, y^{(i)}) = \underset{\theta \in \Theta}{\operatorname{argmin}}\, J(\theta) \tag{2.1.31}$$

Another Interpretation of Linear Regression Suppose $d \ll n$. Assume $\mathbf{X}$ is full-ranked, that is, rank($\mathbf{X}$)=d. Then, the $col(X)$ is a d dimensional subspace of $\mathbb{R}^n$. The projection matrix is $\mathbf{X}(\mathbf{X}^T\mathbf{X})\mathbf{X}^{-1}$, and

$$\hat{y} = \mathbf{X}(\mathbf{X}^T\mathbf{X})\mathbf{X}^{-1}y \tag{2.1.32}$$

for any $y \in \mathbb{R}^n$. Note that, since $d \ll n$, most labels are not in the column space of $\mathbf{X}$, then the projection $\hat{y}$ gives the best approximation of y in the column space of $\mathbf{X}$.