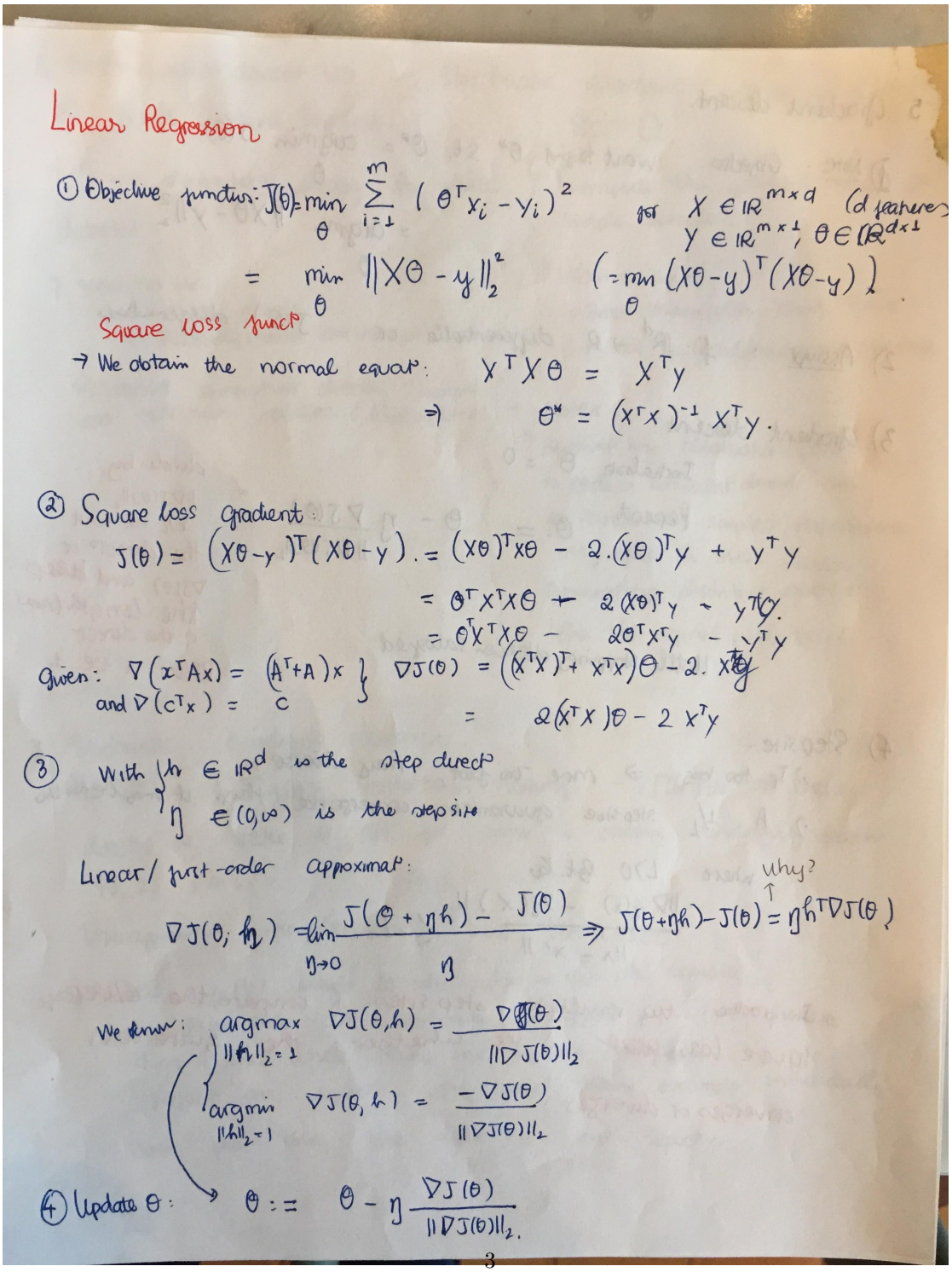
Machine Learning and Computational Statistics
Linear Regression

Nhung Le

Note: This document consists of concepts and exercises related to Linear Regression.

1 Concepts



Linear Regression

① Objective function: $J(\theta) = \min_\theta \sum_{i=1}^{m} (\theta^T x_i - y_i)^2$ for $X \in \mathbb{R}^{m \times d}$ (d features), $y \in \mathbb{R}^{m \times 1}$, $\theta \in \mathbb{R}^{d \times 1}$

$= \min_\theta \|X\theta - y\|_2^2$ $\quad (= \min_\theta (X\theta - y)^T (X\theta - y))$

Square loss funct°

→ We obtain the normal equat°: $X^T X\theta = X^T y$

$\Rightarrow \quad \theta^* = (X^T X)^{-1} X^T y.$

② Square loss gradient:

$J(\theta) = (X\theta - y)^T (X\theta - y) = (X\theta)^T X\theta - 2.(X\theta)^T y + y^T y$

$= \theta^T X^T X\theta + 2(X\theta)^T y + y^T y$

$= \theta^T X^T X\theta - 2\theta^T X^T y - y^T y$

Given: $\nabla(x^T A x) = (A^T + A)x$ and $\nabla(c^T x) = c$ $\Rightarrow \nabla J(\theta) = (X^T X)^T + X^T X)\theta - 2.X^T y$

$= 2(X^T X)\theta - 2X^T y$

③ With $h \in \mathbb{R}^d$ is the step direct°

$\eta \in (0, \infty)$ is the step size

Linear / first-order approximat°:

$\nabla J(\theta; h) = \lim_{\eta \to 0} \frac{J(\theta + \eta h) - J(\theta)}{\eta} \Rightarrow J(\theta + \eta h) - J(\theta) = \eta h^T \nabla J(\theta)$ (why?)

We know: $\underset{\|h\|_2 = 1}{\text{argmax}} \nabla J(\theta, h) = \frac{\nabla J(\theta)}{\|\nabla J(\theta)\|_2}$

$\underset{\|h\|_2 = 1}{\text{argmin}} \nabla J(\theta, h) = \frac{-\nabla J(\theta)}{\|\nabla J(\theta)\|_2}$

④ Update θ: $\theta := \theta - \eta \frac{\nabla J(\theta)}{\|\nabla J(\theta)\|_2}$.

7. Regularization.

7.1 Norm 1 – L_1 – Lasso regularization.

→ Objective function:

$$\hat{w} \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{m} (h_w(x_i) - y_i)^2 + \lambda \|w\|_1$$

? We do the total square loss funct

For the avg square loss,

$$\hat{w} \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{m} (h_w(x_i) - y_i)^2 + \lambda \|w\|_1$$

⇒ the Objectv funct: $\min J(\theta) = \frac{1}{n} (Xw - y)^T (Xw - y) + \lambda \|w\|_1$

$\min$ or $J(\theta) = (Xw - y)^T (Xw - y) + \lambda \|w\|_1$

For w = the minimizer of $J(\theta)$.

→ The Lasso objective funct is not differentiable ($\|w\|_1$ non differentiable) so we can't simply apply gradient descent to find the coef w. ⇒ Use shooting algorith

→ Shooting algorithm / Coordinate descent for Lasso.

→ Obj: At each step we optimize over one component of the unknown parameter vector, fixing all other components.

→ we can find a closed form solut for optimizat over a single component fixing all other components.

1) Lasso properties.

a) λ_{max}:

$$J(w) = \|Xw - y\|_2^2 + \lambda \|w\|_1.$$

→ $J(w)$ is convex.

b) The one-sided directional derivative of $J(w)$:

$$J'(w, v) = \lim_{h \downarrow 0} \frac{J(w + hv) - J(w)}{h}.$$

c) w^* is the minimizer of $J(w)$ ⇔ the directional derivative $J'(w^*; v) \geq 0 \quad \forall v \neq 0$.

Thus $\forall v \neq 0$, $J'(0; v) \geq 0 \Leftrightarrow \lambda \geq C$ for $C = \frac{2|(Xv)^T y|}{\|v\|_1}$

d) $\lambda_{max} = 2\|X^T y\|_\infty$.

λ_{max} is the maximum of the lower bounds of λ.

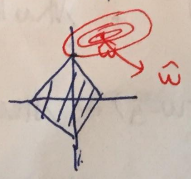
2) Feature correlation.

If X_i and X_j are the same, Lasso would divide the weight arbitrary.

↳ ~~symbols~~ when X_i and X_j are highly correlated, l_1 regularization chooses one w/ larger scale, 0 weight to other

3. Lasso gives sparse solutions

·) l_1 contour $|w_1| + |w_2| = r$ (In 2-dimensions space)



$$\hat{w}_r^* = \underset{w \in \mathbb{R}^2}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

subject to $|w_1| + |w_2| \le r$.

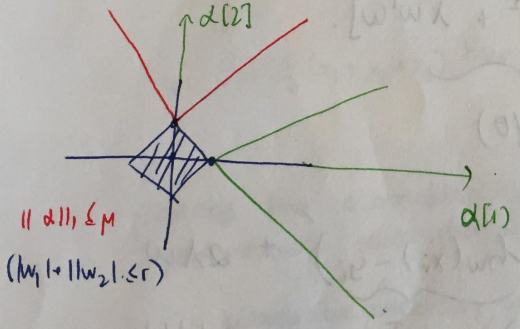
▨ area satisfying $|w_1| + |w_2| \le r$

◎ contour of $\hat{R}_n(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$

·) $\hat{R}_n(w) = \frac{1}{n} \| Xw - y \|^2$, $\hat{R}_n(\hat{w}) = \frac{1}{n} (-y^T X \hat{w} + y^T y)$.

$\hat{R}_n(w)$ is minimized by $\hat{w} = (X^T X)^{-1} X^T y$

for $\hat{R}_n(w) = \frac{1}{n} (w - \hat{w})^T X^T X (w - \hat{w}) + \hat{R}_n(\hat{w})$



·) If X is orthogonal, then $X^T X = I$ and contours are circles

·) Then OLS solution in green/red regions, simples l_1 constrained solution will be at corner

⇒ 1 coef is set to be 0!!