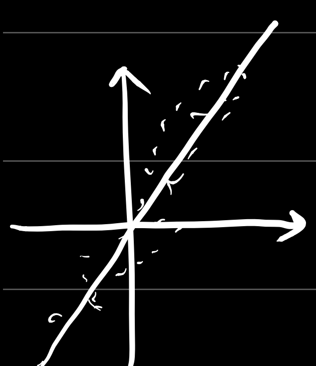
Midterm Review

1 Linear Methods in Regression (Supervised Learning)

Problem Formulation: Given n sample points $x^{(1)}, \cdots, x^{(n)} \in \mathbb{R}^p$ (each has p features) and the corresponding labels $y_1, \cdots, y_n$.
Find coeffi $\beta_1, \cdots, \beta_p$ such that the linear function

$$f(x) = \beta_1 x_1 + \cdots + \beta_p x_p$$

"best approximate" $(x^{(i)}, y_i)$, $i=1,\cdots,n$.



Setting: Write $X = \begin{pmatrix} | & & | \\ x^{(1)} & \cdots & x^{(n)} \\ | & & | \end{pmatrix}$ sample matrix

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{vector of labels}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{vector of coeffi.}$$

(1) Linear Regression:

Idea: choose β by minimizing the sum of the squared distance (mean squared error) over sample pts.

$$\beta^{ls} \overset{\text{def}}{=} \arg\min_{\beta} \| y - X^T \beta \|^2$$

derivation $\Downarrow$

$$\beta^{ls} = (XX^T)^{-1} Xy$$

Related topics:
- β^{ls} tends to overfit if X has "small" singular values.
- SVD

(2) Ridge Regression / ℓ^2-regularization (to reduce overfitting)

Idea: choose β by minimizing the mean squared error while penalizing the ℓ^2-norm of β.

$$\beta^{\text{ridge}} \overset{\text{def}}{=} \arg\min_{\beta} \underbrace{\|y - X^T\beta\|^2 + \lambda \|\beta\|_2^2}_{f(\beta)}$$

where $\lambda > 0$ is a tuning parameter.

derivation: $f(\beta)$ is smooth, convex

take $\frac{\partial f}{\partial \beta} = 0$ to get β^{ridge}

$$\beta^{\text{ridge}} = (XX^T + \lambda I)^{-1} Xy$$

Related topics:
- definition of convex set, convex function, convex optimization.
- local/global minimizer, critical points.

- equivalence of local and global minimizers for convex optimization problem.

(3) Lasso Regression / ℓ^1-regularization (to reduce over fitting)

Idea: similar to Ridge regression but with ℓ^2-norm replaced by ℓ^1-norm.

$$\beta^{lasso} \stackrel{def}{=} \arg\min_{\beta} \underbrace{\|y - X^T\beta\|^2 + \lambda \|\beta\|_1}_{f(\beta)}$$

derivation when $XX^T = I$

$f(\beta)$ is non-smooth, convex

$$\beta_i^{lasso} = S_{\frac{\lambda}{2}}(\beta_i^{ls})$$

where $S_{\frac{\lambda}{2}}(\cdot)$ is the soft thresholding func

Related topics: • def of subgradient, subdifferential

- optimality condition.
- def of the soft thresholding func.

2. Principal Component Analysis (Unsupervised Learning)

Problem Formulation 1: Given n sample pts $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$, find the d-dim subspace S with the maximum projected data variance.

Setting: $X = \begin{pmatrix} | & & | \\ x^{(1)} & \cdots & x^{(n)} \\ | & & | \end{pmatrix}$

The projected data variance along the direction $\vec{u}$

$$= \frac{1}{n} u^T (XX^T) u$$

This is a Rayleigh quotient.

derivation ||

$\Downarrow$

Conclusion: the direction with the ith largest projected data variance is an eigenvector $u^{(i)}$ associated to the ith largest eigenvalue of XX^T

Related topics:

- Rayleigh quotient
- characterization of eigenvalues using Rayleigh quotient
- relation with SVD: if $X = U\Sigma V^T$ is the SVD, then $u^{(i)}$ is the ith column of U.