



Linear Regression

Given  $\mathbf{X} \in \mathbb{R}^p$ ,  $Y \in \mathbb{R}$ , consider linear(affine) prediction rules,  
 $F_{\text{lin}} := \{\mathbf{x} \mapsto \mathbf{x}'\boldsymbol{\beta} + \beta_0 : \boldsymbol{\beta} \in \mathbb{R}^p, \beta_0 \in \mathbb{R}\}$

Empirical risk minimization

**Empirical risk** is the sample average of squared error:  
 $\hat{R}(f) = \hat{\mathbb{E}}_n \ell(f(\mathbf{X}), Y) = \frac{1}{n} \sum_{i=1}^n n(f(\mathbf{X}_i) - Y_i)^2$   
Choose  $\hat{f} := \arg \min_{f \in F_{\text{lin}}} \hat{\mathbb{E}}_n \ell(f(\mathbf{X}), Y)$

Find  $\hat{f} : \mathbf{x} \mapsto \mathbf{x}'\hat{\boldsymbol{\beta}}$ , such that  
 $\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n (\mathbf{X}'_i \boldsymbol{\beta} - Y_i)^2}_{\text{RSS}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2$

where **design matrix**  $\mathbf{X} \in \mathbb{R}^{n \times p}$  and **response vector**  $\mathbf{y} \in \mathbb{R}^n$ .

**Normal equations:**  $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

**Projection Theorem** also leads to normal equations:  
 $(\mathbf{y} - \hat{\mathbf{y}})^{-1}\mathbf{X} = 0 \Leftrightarrow \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \Leftrightarrow \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$

Linear model with additive Gaussian noise

Model the conditional distribution of  $Y$  given  $\mathbf{X} = \mathbf{x}$  as:  
 $P(Y|\mathbf{X} = \mathbf{x}) = \mathcal{N}(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$   
Equivalently:  $Y = \mathbf{x}'\boldsymbol{\beta} + \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \sigma^2)$

**Maximum likelihood** is least square:  
 $L(\boldsymbol{\beta}) = \prod_{i=1}^n p(Y_i|\mathbf{X}_i, \boldsymbol{\beta}) \Leftrightarrow \ell(\boldsymbol{\beta}) = g(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{X}'_i \boldsymbol{\beta})^2$

Fix  $\mathbf{X}$ . Provided  $\mathbb{E}\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$  and  $\text{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}$

**Bayesian analysis:** Treat  $\boldsymbol{\beta}$  as a r.v. with prior distribution  $\mathcal{N}(0, \tau^2\mathbf{I})$ , then compute posterior distribution  $P(\boldsymbol{\beta}|\mathbf{X}, Y)$ .

$P(\boldsymbol{\beta}|\mathbf{X}_1, Y_1, \dots, \mathbf{X}_n, Y_n) \propto P(Y_1, \dots, Y_n|\boldsymbol{\beta}, \mathbf{X}_1, \dots, \mathbf{X}_n)P(\boldsymbol{\beta})$   
 $P(\boldsymbol{\beta}|\mathbf{X}_1, Y_1, \dots, \mathbf{X}_n, Y_n) \propto \exp(-\frac{1}{2}(\sum_{i=1}^n \frac{(Y_i - \mathbf{X}'_i \boldsymbol{\beta})^2}{\sigma^2} + \frac{1}{\tau^2}\|\boldsymbol{\beta}\|^2))$

Linear Regression Regularization

**Trading off bias and variance:** some increase in bias can give a big decrease in variance.

**Subset selection** is like  $L0$  regularization: RSS decreases as the complexity increases because the best fit with a smaller subset is always possible with a larger subset.

**Find a path through subset space:** using cross-validation and forward-stepwise selection or backward-stepwise selection (need  $n > p$ ).

**Ridge regression** is like  $L2$  regularization:  
 $\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} (\sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p p\beta_j^2)$   
 $\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}'\mathbf{y}$

**Lasso regression** is like  $L1$  regularization:  
 $\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} (\sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p p|\beta_j|)$

While ridge regression leads to reduced, but non-zero values of the coefficients, Lasso regression forces some coefficients to be zero.

**Bayesian analysis:** Ridge regression is equivalent to a MAP estimate with a gaussian prior. Lasso regression is equivalent to a MAP estimate with a Laplace prior.

Logistic Regression

Model **log odds**  $(\log p/(1 - p))$  as an affine function of  $\mathbf{x}$ .

$P(Y = 1|\mathbf{x}) = \frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x})}$  Given data  
 $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n) \in \mathbb{R}^p \times \{0, 1\}$ , estimate  $\boldsymbol{\beta}$  with maximum likelihood.

**Log likelihood:**  
 $\ell(\boldsymbol{\beta}) = \sum_{i=1}^n y_i \log \mu_i(\boldsymbol{\beta}) + (1 - y_i) \log(1 - \mu_i(\boldsymbol{\beta}))$ ,  
where  $\mu_i(\boldsymbol{\beta}) = P(Y = 1|\mathbf{X} = \mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{1+\exp(-\boldsymbol{\beta}'\mathbf{x}_i)}$

$\nabla_{\boldsymbol{\beta}} \mu_i(\boldsymbol{\beta}) = \mu_i(\boldsymbol{\beta})(1 - \mu_i(\boldsymbol{\beta}))\mathbf{x}_i$   
 $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mu_i(\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{X}'(\mathbf{y} - \boldsymbol{\mu})$   
 $\nabla_{\boldsymbol{\beta}}^2 \ell(\boldsymbol{\beta}) = \sum_{i=1}^n -\mu_i(\boldsymbol{\beta})(1 - \mu_i(\boldsymbol{\beta}))\mathbf{x}_i \mathbf{x}'_i = -\mathbf{X}'\text{diag}(\boldsymbol{\mu}(1 - \boldsymbol{\mu}))\mathbf{X}$   
 $\hat{\boldsymbol{\beta}}^{\text{ml}}$  solves:  $\sum_{i=1}^n y_i \mathbf{x}_i = \sum_{i=1}^n \mu_i \boldsymbol{\beta} \mathbf{x}_i$

**Gradient ascent:**  
 $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \eta \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^{(t)}) : O(np)/\text{step}$   
**Stochastic gradient ascent:**  
 $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \eta (y_{i_t} - \mu_{i_t}(\boldsymbol{\beta}^{(t)}))\mathbf{x}_{i_t} : O(p)/\text{step}$   
**Newton-Raphson method:**  
 $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - [\nabla_{\boldsymbol{\beta}}^2 \ell(\boldsymbol{\beta}^{(t)})]^{-1} \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^{(t)})$

**Newton's method for root finding:**  $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

**Prediction**  $\hat{p}(y|\mathbf{x}) = \begin{cases} P(Y = 1|\mathbf{x}), & \text{if } y = 1 \\ P(Y = -1|\mathbf{x}), & \text{if } y = -1 \end{cases}$

**Log loss (Binomial Deviance):**  $\ell_{\log}(\hat{p}(\cdot|\mathbf{x}), y) = -\log(\hat{p}(y|\mathbf{x}))$   
**Minimize:**  $\frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i \boldsymbol{\beta}'\mathbf{x}_i))$

Linear Discriminant Analysis

**Linear discriminant functions:**  
 $\delta_k(\mathbf{x}) = \boldsymbol{\mu}'_k \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}'_k \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k$

**Estimate with Maximum likelihood:**  
 $\pi_k = P(Y = k) \Leftrightarrow \hat{\pi}_k = \frac{n_k}{n}$   
 $\boldsymbol{\mu}_k = \mathbb{E}[\mathbf{X}|Y = k] \Leftrightarrow \hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i: y_i = k} \mathbf{x}_i$   
 $\boldsymbol{\Sigma} = \text{Var}[\mathbf{X}|Y = k] \Leftrightarrow \hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_k \sum_i i : y_i = k (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)'$

SVM with Convex Optimization

**Lagrangian:** rewrite constraint as penalties for a convex optimization problem such that  $L(x, \lambda) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x)$ .

**Weak duality:**  $p^* = \min_x \max_{\lambda \geq 0} L(x, \lambda) \geq \max_{\lambda \geq 0} \min_x L(x, \lambda) = d^*$   

$\underbrace{\hspace{10em}}$   
primal

$\underbrace{\hspace{10em}}$   
dual

**Strong duality:**  
if there is a saddle point  $(x^*, \lambda^*)$  such that for all  $x$  and  $\lambda \geq 0$ ,  $L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*)$ , then primal and dual have the same value ( $p^* = d^*$ ).

**Karush-Kuhn-Tucker optimality conditions:**  
Primal feasibility:  $f_i(x) \leq 0$ ; Dual feasibility:  $\lambda_i \geq 0$   
Complementary slackness:  $\lambda_i f_i(x) = 0$   
Stationarity:  $\nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) = 0$

**Hard margin SVM:**  
 $L(\boldsymbol{\theta}, \alpha) = \frac{1}{2} \|\boldsymbol{\theta}\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i \boldsymbol{\theta}'\mathbf{x}_i)$   
 $g(\alpha) = \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \alpha)$   
setting  $\boldsymbol{\theta}^* = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$ ,  
 $g(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}'_i \mathbf{x}_j$

**Hard margein SVM dual problem:**  
 $\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}'_i \mathbf{x}_j$ , s.t.  $\alpha_i \geq 0 \ (i = 1, \dots, n)$ .  
 $\min_{\alpha} \frac{1}{2} \boldsymbol{\alpha}' \text{diag}(\mathbf{y}) \mathbf{K} \text{diag}(\mathbf{y}) \boldsymbol{\alpha} - \boldsymbol{\alpha}' \mathbf{1}$ , s.t.  $\boldsymbol{\alpha} \geq \mathbf{0}$ .

**Soft margin SVM:**  
 $L(\boldsymbol{\theta}, \xi, \alpha, \lambda) = \frac{1}{2} \|\boldsymbol{\theta}\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i (1 - y_i \boldsymbol{\theta}'\mathbf{x}_i - \xi_i) - \sum_{i=1}^n \lambda_i \xi_i$

**Soft margein SVM dual problem:**  
 $\min_{\alpha} \frac{1}{2} \boldsymbol{\alpha}' \text{diag}(\mathbf{y}) \mathbf{K} \text{diag}(\mathbf{y}) \boldsymbol{\alpha} - \boldsymbol{\alpha}' \mathbf{1}$ , s.t.  $\mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{C}{n}$ .