CS189 - HOW TO LEARN MACHINES | • Perceptron: $\theta^{t+1} = \theta^t + \underbrace{y_i x_i}_{\text{some misclassified pair}}$

• SVMs:

hard margin: $\min_\theta \|\theta\|_2^2 : y_i \theta^T x_i \geq 1 \quad i=1,\dots,n$ (LS) → margin width $\frac{1}{\|\theta\|_2}$

soft margin: $\min_\theta \frac{1}{2}\|\theta\|_2^2 + C\cdot\sum_{i=1}^n (1-y_i\theta^T x_i)_+$ (QP)

• Bayes decision: $f^*(x) = \begin{cases} 1 & P(Y=1|X) > P(Y=-1|X) \\ -1 & \text{o.w.} \end{cases}$

• Generative models use class conditionals $P(X|Y)$

• Discriminative models use $P(Y|X)$

MLE: $L = P(X \dots X_n|\theta)$

MAP: $L = P(X_1 \dots X_n|\theta)P(\theta)$

equivalent when $P(\theta)$ uniform

Gaussian generative: $P(X|\omega_i) = \mathcal{N}(\mu_i, \Sigma)$ → Logistic discriminative: $P(\omega_i|x) = (1+\exp(-\theta^T x - \theta_0))^{-1}$

Two class case: $\theta = \frac{\mu_1 - \mu_0}{\sigma^2}$, $\theta_0 = \frac{\mu_0^2 - \mu_1^2}{2\sigma^2} - \log\left(\frac{P(0)}{P(1)}\right)$

Using kernel for SVM, classify test point x by $\theta \cdot \phi(x) = \sum_j \alpha_j y_j (\phi(x_j)\cdot\phi(x)) = \sum_j \alpha_j y_j K(x_j, x)$ where $\alpha_j \neq 0$ only for support vectors. (points for which $y_i \theta^T x_i = 1$)

Multivariate Gaussian; $x \in \mathbb{R}^d$:

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}$$

properties: • If X, Y independent, d-dimensional Gaussians, $X+Y \sim \mathcal{N}(\mu_x+\mu_y, \Sigma_x+\Sigma_y)$

• If $X \sim \mathcal{N}(\mu, \Sigma)$, $Y = AX+b$, $Y \sim \mathcal{N}(A\mu+b, A\Sigma A^T)$

SVM hard margin dual:

$$\max_\alpha \sum_{i=n}^n \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j : \alpha_i \geq 0$$

Soft margin dual: corresponding primal:

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j : 0 \leq \alpha_i \leq \frac{C}{n} \to \min_{\theta,\xi} \frac{1}{2}\|\theta\|_2^2 + \frac{C}{n}\sum_i \xi_i : 1 - y_i\theta^T x_i - \xi_i \leq 0,\ \xi_i \geq 0$$

2-class ($y \in \{0,1\}$) logistic regression: Assumes $P(Y=1|X) = (1+e^{-\beta^T x})^{-1}$

log-likelihood $\ell(\beta) = \sum_{i=1}^n y_i \log \mu_i(\beta) + (1-y_i)\log(1-\mu_i(\beta))$, $\mu_i(\beta) = \frac{1}{1+e^{-\beta^T x_i}} = P(Y=1|X=x_i,\beta)$

NB: $\nabla_\beta \mu_i(\beta) = \mu_i(\beta)(1-\mu_i(\beta))x_i \to \nabla_\beta \ell(\beta) = \sum_{i=1}^n (y_i - \mu_i(\beta))x_i$

$\nabla_\beta^2 \ell(\beta) = \sum_{i=1}^n -\mu_i(\beta)(1-\mu_i(\beta))x_i x_i^T$

Newton's method:

root finding: $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

extrema: $x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$

Gradient ascent/descent: $\beta^{(t+1)} = \beta^{(t)} \pm \eta\nabla_\beta \ell(\beta^{(t)})$

Newton-Raphson update: $\beta^{(t+1)} = \beta^{(t)} - [\nabla_\beta^2 \ell(\beta^{(t)})]^{-1}\nabla_\beta \ell(\beta^{(t)})$

With Gaussian class conditional/logistic posterior, linear discriminant analysis uses function

$\delta_k(x) = \mu_k^T\Sigma^{-1}x - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$, can estimate $\hat{\mu}_k, \hat{\pi}_k, \hat{\Sigma}$ from sample mean/prior/covariance

pick class k which maximizes $\delta_k(x)$

Lagrangian duality

$p^* = \min_x f_0(x) : f_i(x) \leq 0 \to L(x,\lambda) = f_0(x) + \sum_i \lambda_i f_i(x)$

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

↑ equal when strongly dual

Strong duality KKT optimality conditions:

$f_i(x) \leq 0$, $\lambda_i \geq 0$ feasible

$\lambda_i f_i(x) = 0$ complementary slackness

$\nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) = 0$ stationarity

For SVMs, duality means

$\theta^* = \sum_i \alpha_i^* y_i x_i$, $\alpha_i^* = 0$ if x_i^* is not a support vector

at optimum, $\xi_i^* = (1-y_i\theta^T x_i)_+$

$\alpha_i^*(1-y_i x_i^T\theta^* - \xi_i^*) = 0$

$\lambda_i^*\xi_i^* = 0$

where $\alpha_i + \lambda_i = \frac{C}{n}$

or

$\alpha_i(y_i(w^T x_i + b) - 1) = 0$

PK $K = (x^T y + c)^d$

Quadratic kernel features: $\phi(x) = [c, x_1^2, \dots, x_n^2, \dots \sqrt{2}x_1x_2 \dots \sqrt{2}x_1x_n, \sqrt{2}x_2x_3, \dots \sqrt{2}x_{n-1}x_n, \sqrt{2c}\,x_1 \dots \sqrt{2c}\,x_n]^T$

Method of Lagrange multipliers

min/max $f_0(x)$ s.t. $f_1(x) = 0$

$L(x, \lambda) = f_0(x) + \lambda f_1(x)$

take P.D.'s, set to 0