

Probability

Bayes Theorem:  $P(Y = \pm 1|X) = \frac{P(X|Y=\pm 1)P(Y=\pm 1)}{P(X|Y=+1)P(Y=+1)+P(X|Y=-1)P(Y=-1)}$

Perceptron

$f(\mathbf{x}) = \boldsymbol{\theta} \cdot \mathbf{x} + \theta_0 = \sum_{i=1}^d \theta_i x_i + \theta_0, \hat{y} = \begin{cases} 1, & \text{if } f(x) \geq 0 \\ -1, & \text{if } f(x) < 0 \end{cases}$

Decision boundary, a hyperplane in  $\mathbb{R}^d$ :  
 $H = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = 0\} = \{\mathbf{x} \in \mathbb{R}^d : \boldsymbol{\theta} \cdot \mathbf{x} + \theta_0 = 0\}$

$\boldsymbol{\theta}$  is the normal of the hyperplane,  
 $\theta_0$  is the offset of the hyperplane from origin,  
 $-\frac{\theta_0}{\|\boldsymbol{\theta}\|}$  is the signed distance from the origin to hyperplane.

Perceptron algorithm,  
Input:  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \{\pm 1\}$   
while some  $y_i \neq \text{sign}(\boldsymbol{\theta} \cdot \mathbf{x}_i)$   
    pick some misclassified  $(\mathbf{x}_i, y_i)$   
     $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + y_i \mathbf{x}_i$

Given a linearly separable data, perceptron algorithm will take no more than  $\frac{R^2}{\gamma^2}$  updates to converge, where  $R = \max_i \|\mathbf{x}_i\|$  is the radius of the data,  $\gamma = \min_i \frac{y_i(\boldsymbol{\theta} \cdot \mathbf{x}_i)}{\|\boldsymbol{\theta}\|}$  is the margin.  
Also,  $\frac{\boldsymbol{\theta} \cdot \mathbf{x}}{\|\boldsymbol{\theta}\|}$  is the signed distance from H to  $\mathbf{x}$  in the direction  $\boldsymbol{\theta}$ .

$\boldsymbol{\theta} = \sum_i \alpha_i y_i \mathbf{x}_i$ , thus any inner product space will work, this is a kernel.

Gradient descent view of perceptron, minimize margin cost function  $J(\boldsymbol{\theta}) = \sum_i (-y_i(\boldsymbol{\theta} \cdot \mathbf{x}_i))_+$  with  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla J(\boldsymbol{\theta})$

Support Vector Machine

Hard margin SVM,  
 $\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|^2$ , such that  $y_i \boldsymbol{\theta} \cdot \mathbf{x}_i \geq 1 (i = 1, \dots, n)$   
Soft margin SVM,  
 $\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n (1 - y_i \boldsymbol{\theta} \cdot \mathbf{x}_i)_+$

Regularization and SVMs: Simulated data with many features  $\phi(\mathbf{x})$ ; C controls trade-off between margin  $1/\|\boldsymbol{\theta}\|$  and fit to data; Large C: focus on fit to data (small margin is ok). More overfitting. Small C: focus on large margin, less tendency to overfit. Overfitting increases with: less data, more features.

$\boldsymbol{\theta} = \sum_j \alpha_j y_j \mathbf{x}_j, \alpha_j \neq 0$  only for support vectors.

$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ , K is called a kernel.  
Solve  $\alpha_j$  to determine  $\sum_j \alpha_j y_j \phi(\mathbf{x}_j)$ .  
Compute the classifier for a test point  $\mathbf{x}$  via  
 $\boldsymbol{\theta} \cdot \phi(\mathbf{x}) = \sum_j \alpha_j y_j K(\mathbf{x}_j, \mathbf{x})$

degree-m polynomial kernel:  $K_m(\mathbf{x}, \tilde{\mathbf{x}}) = (1 + \mathbf{x} \cdot \tilde{\mathbf{x}})^m$   
radial basis function kernel:  $K_{rbf}(\mathbf{x}, \tilde{\mathbf{x}}) = \exp(-\gamma \|\mathbf{x} - \tilde{\mathbf{x}}\|^2)$

Decision Theory

Loss function:  $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ , and  $l(\hat{y}, y)$  is the cost of predicting  $\hat{y}$  when the outcome is  $y$ .

Assume  $(\mathbf{X}, \mathbf{Y})$  are chosen i.i.d according to some probability distribution on  $\mathcal{X} \times \mathcal{Y}$ . Risk is misclassification probability:  
 $R(f) = \mathbb{E}l(f(\mathbf{X}), \mathbf{Y}) = Pr(f(\mathbf{X}) \neq \mathbf{Y})$

Bayes Decision Rule is  
 $f^*(x) = \begin{cases} 1, & \text{if } P(\mathbf{Y} = 1|x) > P(\mathbf{Y} = -1|x) \\ -1, & \text{otherwise.} \end{cases}$ ,  
and the optimal risk (Bayes risk)  $R^* = \inf_f R(f) = R(f^*)$

Excess risk is for any  $f : \mathcal{X} \rightarrow \{-1, +1\}$ ,  
 $R(f) - R^* = \mathbb{E}(1[f(x) \neq f^*(x)]|2P(\mathbf{Y} = +1|\mathbf{X}) - 1|)$

Risk in Regression is expected squared error:  
 $R(f) = \mathbb{E}l(f(\mathbf{X}), \mathbf{Y}) = \mathbb{E}\mathbb{E}[f(\mathbf{X}) - \mathbf{Y}^2|\mathbf{X}]$

Bias-variance decomposition:  
 $R(f) = \mathbb{E}[\underbrace{(f(\mathbf{X}) - \mathbb{E}[\mathbf{Y}|\mathbf{X}])^2}_{\text{bias}^2} + \underbrace{\mathbb{E}[(\mathbb{E}[\mathbf{Y}|\mathbf{X}] - \mathbf{Y})^2]}_{\text{variance}}]$

Generative and Discriminative

Discriminative models:  $P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X})P(\mathbf{Y}|\mathbf{X})$ .  
Estimate  $P(\mathbf{Y}|\mathbf{X})$ , then pretend out estimate  $\hat{P}(\mathbf{Y}|\mathbf{X})$  is the actual  $P(\mathbf{Y}|\mathbf{X})$  and plug in bayes rule expression.

Generative model:  $P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{Y})P(\mathbf{X}|\mathbf{Y})$ .  
Estimate  $P(\mathbf{Y})$  and  $P(\mathbf{X}|\mathbf{Y})$ , then use bayes theorem to calculate  $P(\mathbf{Y}|\mathbf{X})$  and use discriminative model.

Estimation

Method of moments: Match moments of the distribution to momemnts measured in the data.

Maximum likelihood: Choose parameter so that the distribution it defines gives the observed data the highest probability (likelihood).

Maximum log likelihood: Log of maximum likelihood, equivalent to maximum likelihood since log is monotonically increase; it is useful since it can change  $\prod$  to  $\sum$ .

Penalized maximum likelihood: Add a penalty term in the maximum (log) likelihood equation; treat the penalty term as some imaginary data points crafted for desired probability.

Bayesian estimate: Treat parameter as a random variable, then update based on observed value (data).  
Prior:  $\pi(p) = 1$ ,  
Posterior:  $P(p|\mathbf{X}_1 = 1) = P(\mathbf{X}_1 = 1|p)\pi(p)/\int P(X_1 = 1|q)d\pi(q)$

Maximum a posterior probability: the mode of the posterior.  
If uniform prior, MAP is MLE; if not uniform prior, MAP is Penalized MLE.

Gaussian maximum likelihood estimation:  
 $\mu = \frac{1}{n} \sum_{i=1}^n x_i, \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$