1 PAC Learning

Empirical error: $\hat{\mathcal{R}}_n(c) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{c(x_i)\neq y\}}$
Expected error: $\mathcal{R}(c) = P\{c(x) \neq y\}$
ERM: $\hat{c}_n^* = \arg\min_{c\in\mathcal{C}} \hat{\mathcal{R}}_n(c)$
opt: $c^* \in \min_{c\in\mathcal{C}} \mathcal{R}(c)$, $|\mathcal{C}|$ finite
Generalization error: $\mathcal{R}(\hat{c}_n^*) = P\{\hat{c}_n^*(x) \neq y\}$
VC ineq.: $\mathcal{R}(\hat{c}_n^*) - \inf_{c\in\mathcal{C}} \mathcal{R}(c) \leq 2\sup_{c\in\mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)|$
$P\{\mathcal{R}(\hat{c}_n^*) - \mathcal{R}(c^*) > \epsilon\} \leq P\{\sup_{c\in\mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| > \frac{\epsilon}{2}\}$
$\leq 2|\mathcal{C}|exp(-2n\epsilon^2/4) \leq 8s(\mathcal{A},n)exp(-n\epsilon^2/32)$
and $s(\mathcal{A},n) \leq n^{\mathcal{V}_{\mathcal{A}}}$
Markov ineq: $P\{X \geq \epsilon\} \leq \frac{\mathbb{E}[X]}{\epsilon}$ (for nonneg. X)
Boole's inequality: $P(\bigcup_i A_i) \leq \sum_i P(A_i)$
Hoeffding's lemma: $\mathbb{E}[e^{sX}] \leq exp(\frac{1}{8}s^2(b-a)^2)$
where $\mathbb{E}[X] = 0$, $P(X \in [a,b]) = 1$
Hoeffding's: $P\{S_n - \mathbb{E}[S_n] \geq t\} \leq exp(-\frac{2t^2}{\sum_i (b_i - a_i)^2})$
Normalized: $P\{\widetilde{S}_n - \mathbb{E}[\widetilde{S}_n] \geq \epsilon\} \leq exp(-\frac{2n^2\epsilon^2}{\sum_i (b_i - a_i)^2})$
Error bound: $P\{\sup_{c\in\mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| > \epsilon\} \leq 2|\mathcal{C}|exp(-2n\epsilon^2)$
The $\mathcal{VC}$ dimension of a model f is the maximum number of points that can be arranged so that f shatters them.

2 Nonparametric Bayesian methods

$Dir(x|\alpha) = \frac{1}{B(\alpha)}\prod_{k=1}^{n} x_k^{a_k - 1}$, $B(\alpha) = \frac{\prod_{k=1}^{n}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{n}\alpha_k)}$
$\mathbb{E}[1] = \sum_{i=1}^{N} \frac{\alpha}{\alpha + i} \sim (\alpha log(N))$
de Finetti: $p(X_1, ..., X_n) = \int (\prod_{i=1}^{n} p(x_i|G)) dP(G)$

$$p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}, \alpha, \boldsymbol{\mu}) = \begin{cases} \frac{N_{k,-i}}{\alpha+N-1} p(x_i|\boldsymbol{x}_{-i,k}, \boldsymbol{\mu}) & \exists k \\ \frac{\alpha}{\alpha+N-1} p(x_i|\boldsymbol{\mu}) & \text{otherwise} \end{cases}$$

DP generative model:
- Centers of the clusters: $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0)$
- Prob.s of clusters: $\rho = (\rho_1, \rho_2) \sim GEM(\alpha)$
- Assignments to clusters: $z_i \sim Categorical(\rho)$
- Coordinates of data points: $\mathcal{N}(\mu_{z_i}, \sigma)$

3 Generative Methods

Naive Bayes

All features independent.
$P(y|x) = \frac{1}{Z}P(y)P(x|y), Z = \sum_y P(y)P(x|y)$
$y = \arg\max_{y'} P(y'|x) = \arg\max_{y'} \hat{P}(y')\prod_{i=1}^{d} \hat{P}(x_i|y')$

Discriminant Function

$f(x) = \log(\frac{P(y=1|x)}{P(y==1|x)}), y = sign(f(x))$

4 Neural Networks

Learning features

Parameterize the feature maps and optimize over the parameters:
$w^* = \operatorname{argmin}_{w,\Theta} \sum_{i=1}^{n} l(y_i, \sum_{j=1}^{m} w_j \Phi(x_i, \Theta_j))$

Reformulating the perceptron

Ansatz: $w = \sum_{j=1}^{n} \alpha_j y_j x_j$
$\min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} \max[0, -y_i w^T x_i]$
$= \min_{\alpha_{1:n}} \sum_{i=1}^{n} \max[0, -y_i(\sum_{j=1}^{n} \alpha_j y_j x_j)^T x_i]$
$= \min_{\alpha_{1:n}} \sum_{i=1}^{n} \max[0, -\sum_{j=1}^{n} \alpha_j y_i y_j x_i^T x_j]$

Kernelized Perceptron

1. Initialize $\alpha_1 = ... = \alpha_n = 0$
2. For t do
Pick data $(x_i, y_i) \in_{u.a.r} D$
Predict $\hat{y} = sign(\sum_{j=1}^{n} \alpha_j y_j k(x_j, x_i))$
If $\hat{y} \neq y_i$ set $\alpha_i = \alpha_i + \eta_t$