Basis

Inner Product: $\langle \vec{x}, \vec{z} \rangle = \vec{x}^T \vec{z} = \sum_{i=1}^{d} x_i z_i$

Euclidean Norm: $\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle}$

Hadamar Product. $A \circ B = \begin{pmatrix} a_{11}b_{11} & \cdots & a_{1n}b_{1n} \\ \vdots & & \vdots \\ a_{m1}b_{m1} & \cdots & a_{mn}b_{mn} \end{pmatrix}$

Kronecker Product: $A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}$

Positive Semi-definite Matrix A

$\Leftrightarrow \forall \vec{v} \in \mathbb{R}^n : \vec{v}^T A \vec{v} \geq 0$

$\Leftrightarrow \forall \lambda_i : \lambda_i \geq 0$

Properties:
1. λ_i of $A \otimes B$ are the product of λ_i of A & B
2. $A \circ B$ is a principal submatrix of $A \otimes B$
3. $A \otimes B$ is positive semi-definite $\Leftrightarrow$ $A \circ B$ is as well

Moore-Penrose Pseudoinverse: $A^+ = (A^T A)^{-1} A^T$

Inner Product Space: $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$
1. $\langle x, y \rangle = \langle y, x \rangle$
2. $\langle x, x \rangle \geq 0$
3. $\langle \lambda_1 x + \lambda_2 y, z \rangle = \lambda_1 \langle x, z \rangle + \lambda_2 \langle y, z \rangle$

Computational Learning Theory

Risk: $R[f] = \int_{X \times Y} V(f(x), y)\, dP(x, y)$

Empirical Risk: $R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i)$

Regularization: Limit capacity

Lemma:

$$R[f] \leq R_{emp}[f] + \sqrt{\frac{VC_{dim}(\mathcal{H})\left(\ln \frac{2m}{VC_{dim}(\mathcal{H})} + 1\right) - \ln \frac{\delta}{4}}{m}}$$

Tikhonov Regularization:

$$\operatorname*{argmin}_{f \in \mathcal{H}} R_{emp}[f] + \lambda \Omega[f]$$

Version space: $VS_{\mathcal{H},E} = \{h \in \mathcal{H} : h \text{ is consistent with } h\}$

General Boundary G: $g \in G \Leftrightarrow g$ is consistent $\wedge\ \nexists_{g' \in \mathcal{H}} : g' < g \wedge g'$ is consistent

Specific Boundary S: $s \in S \Leftrightarrow s$ is consistent $\wedge\ \nexists_{s' \in \mathcal{H}} : s' > s \wedge s'$ is consistent

Consistency:
1. h is complete: $E^+ \subseteq h$
2. h is correct: $E^- \cap h = \emptyset$

Generality: $h_1 \leq h_2$: h_1 is more general than h_2

PAC-learnable: $P(error(h) \leq \varepsilon) \geq 1 - \delta$

Efficient PAC-learnable: Alg. runs in $\frac{1}{\varepsilon}, \frac{1}{\delta}, n, size(c)$

Finite concept classes:

Theorem: C is PAC-learnable, if we draw m samples and find a consistent hypothesis with:

$$m \geq \frac{1}{\varepsilon}\left(\ln|\mathcal{H}| + \ln \frac{1}{\delta}\right)$$

Corollary:
1. Theorem holds
2. m is polynomial in n
3. Alg. runs polynomial in m and n

$\Rightarrow C$ is efficiently PAC-learnable

Lemma: $VC_{dim}(C) \leq \log_2 |C|$

Results:
1. k-CNF_n / k-DNF_n is efficiently PAC-learnable
2. Consistency problem for k-term DNF_n is NP-complete
3. 3-term-$DNF_n \leq 3$-CNF_n
4. 3-term DNF_n is efficiently PAC-learnable using 3-CNF_n
5. $MONOM_n$ is efficiently PAC-learnable
6. Concept class of axis-alligned rectangles is eff. PAC-learn.
7. Concept class of symmetric boolean func is eff. PAC-learn.
8. $VC_{dim} = d+1$ for set of half-spaces in $\mathbb{R}^d$
9. $VC_{dim} = 3$ for circles in $\mathbb{R}^2$
10. $VC_{dim} = 2d+1$ for d-gons in $\mathbb{R}^2$

Infinite concept classes:

Family of subsets realized by C: $\Pi_C(S) = \{c \cap S : c \in C\}$

Number of realizable dichotomies: $\Pi_C(m) = \max\{|\Pi_C(S)| : S \subseteq X, |S| = m\}$

Shattering: $S \subseteq X$ is shattered by C, if $\Pi_C(S) = 2^S$

VC-Dimension d: Largest d, s.t. $\exists_{S \subseteq X} : |\Pi_C(S)| = |2^S| = d$

Error region: $\Delta(C) = \{h \Delta c : h \in C\}$

ε-Net: $\Delta_\varepsilon(C) = \{r \in \Delta(C) \cdot P_{r_{x \in D}}[x \in r] \geq \varepsilon\}$

Theorem (BEHW): C is PAC-learnable, if we draw m samples and find a consistent hypothesis with:

$$m \geq \frac{4}{\varepsilon} \log_2 \frac{2}{\delta} + \frac{8 \cdot VC_{dim}(C)}{\varepsilon} \log_2 \frac{13}{\varepsilon}$$
$$= O\left(\frac{1}{\varepsilon}\left(\log \frac{1}{\delta} + VC_{dim}(C) \log \frac{1}{\varepsilon}\right)\right)$$

Corollary:
1. Theorem holds
2. $VC_{dim}(C)$ is polynomial in n
3. Alg. runs polynomial in m and n

$\Rightarrow C$ is efficiently PAC-learnable

Theorem (PSSVC): For $d = VC_{dim}(C) > 0$:
1. $\Pi_C(m) = 2^m$ if $m \leq d$
2. $\Pi_C(m) \leq \left(\frac{me}{d}\right)^d$ else

Lemma: $C_1 \subseteq C_2 \Rightarrow VC_{dim}(C_1) \leq VC_{dim}(C_2)$

Lemma: For $\overline{C} = \{X \setminus c : c \in C\}$: $VC_{dim}(\overline{C}) = VC_{dim}(C)$

Radon's Theorem: Any set S with $d+2$ points in $\mathbb{R}^d$ can be partitioned, s.t.: $Conv(S_1) \cap Conv(S_2) \neq \emptyset$

Mangasarian Theorem: Given: 2 disjoint linearly separable subsets of $\mathbb{R}^d$

Then: Separating hyperplane can be found polynomial in d and $|S_1 \cup S_2|$

Optimization Theory

Constrained opt. problem: $\min_{\vec{w}} f(\vec{w})$ s.t. $g_i(\vec{w}) \leq 0$

Generalized Lagrangian: $L(\vec{w}, \vec{\alpha}) = f(\vec{w}) + \sum_{i=1}^{m} \alpha_i g_i(\vec{w})$

Dual opt. problem: $\max_{\vec{\alpha} \geq 0} \underbrace{\min_{\vec{w}} L(\vec{w}, \vec{\alpha})}_{L_D(\vec{\alpha})}$

Decision Tree Learning

Entropy: $H(S) = \sum_{i}^{n} p_i \log_2 \frac{1}{p_i}$

Lemma: $H(S) \leq \log_2(n)$

Cond. Entropy: $H(S|A) = \sum_{j=1}^{m} P(A = v_j) \cdot H(S_j)$

Lemma: $H(S|A) \leq H(S)$

Information Gain: $H(S) - H(S|A)$

Gini Coefficient: $Gini(S) = 1 - (p_\oplus^2 + p_\ominus^2)$

Gini Split: $Gini_{split}(S, A) = \sum_{v \in values(A)} \frac{|S_v|}{|S|} Gini(S_v)$

Gini Gain: $Gain_{Gini}(S, A) = Gini(S) - Gini_{split}(S, A)$

Kernel Methods

Kernel: $k(x,y) = \langle \phi(x), \phi(y) \rangle$

Gram/Kernel Matrix $K_{X,k}$: $(K_{X,k})_{ij} = k(x_i, x_j)$

Finite Domain: k is kernel

$\Leftrightarrow K_{X,k}$ is positive semi-definite

Infinite Domain: (Mercer's Theorem)

k is kernel $\Leftrightarrow \forall Y \subseteq X: K_{Y,k}$ is pos. semi-definite

Cover's Theorem: Number of lin. separable dichotomies of n points in $\mathbb{R}^d$

$C(n,d) = 2\sum_{k=0}^{d} \binom{n-1}{k}$

Prob. that a rand. dich. is lin. sep.: $P(n,d) = \frac{1}{2^n} C(n,d)$

Phase transition: $n = 2(d+1)$

Kernels:

1. $k(x,y) = f(x)f(y)$
2. $k(\vec{x},\vec{y}) = \vec{x}^T A \vec{y}$
3. $k(x,y) = \alpha k_1(x,y) + \beta k_2(x,y)$
4. $k(x,y) = k_1(x,y)k_2(x,y)$
5. $k(x,y) = p(k_1(x,y))$
6. $k(x,y) = f(k_1(x,y))$
7. $k(x,y) = e^{k_1(x,y)}$
8. $k(\vec{x},\vec{y}) = \exp\left(-\frac{\|\vec{x}-\vec{y}\|^2}{2\sigma^2}\right)$
9. $k(\vec{x},\vec{y}) = (\vec{x}^T\vec{y} + c)^m$

Ridge Regression:

least squares with L2-Tikhonov regularization

Optimization function:

$\min_{w} R_\lambda(\vec{w}, S) = \min_{w} \|X\vec{w} - \vec{y}\|^2 + \lambda\|\vec{w}\|^2$

Gradient: $\nabla_{\vec{w}} R_\lambda(\vec{w}, S) = 2X^TX\vec{w} - 2X^T\vec{y} + 2\lambda\vec{w}$

Primal solution: $\vec{w} = (X^TX + \lambda I_d)^{-1} X^T\vec{y}$

Regression function: $f(\vec{x}) = \vec{x}^T\vec{w}$

Dual solution: $\vec{\alpha} = (XX^T + \lambda I_n)^{-1}\vec{y}$

Regression function: $f(\vec{x}) = \sum_{i=1}^{n} \alpha_i \langle \vec{x}, \vec{x}_i \rangle$

Lasso Regression

least squares with L2-Tikhonov regularization

Support Vector Machines

Hyperplane: $\langle \vec{w}, \vec{x} \rangle + b = 0$

Margin: $\gamma = \frac{2}{\|\vec{w}\|}$

Decision function: $\mathrm{sign}(\langle \vec{w}, \vec{x} \rangle + b)$

γ-Shattering: All dichotomies are realizable through a hyperplane with margin γ

Theorem: For any γ and any $S \subset \{\vec{x} \in \mathbb{R}^d : \|\vec{x}\| \le R\}$

$|S| \le \min\left(\left(\frac{R}{\gamma}\right)^2, d\right) + 1$

Support vectors: $\langle \vec{w}, \vec{x} \rangle + b = \{1, -1\}$

$\alpha_i > 0$

Hard Margin SVM

Primal form:

$\max_{\vec{w},b} \frac{2}{\|\vec{w}\|}$ with $y_i(\langle \vec{w}, \vec{x}_i \rangle + b) \ge 1$

$\Leftrightarrow \min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|^2$ with $-(y_i(\langle \vec{w}, \vec{x}_i \rangle + b) - 1) \le 0$

Dual form:

$\max_{\vec{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \langle \vec{x}_i, \vec{x}_j \rangle$

with $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $\alpha_i \ge 0$

Soft Margin SVM

Soft constraints: $y_i(\langle \vec{w}, \vec{x}_i \rangle + b) \ge 1 - \xi_i$

$\xi_i = 0$: correct

$0 < \xi_i \le 1$: within margin, but correct side

$\xi_i > 1$: wrong side

Primal form: $\min_{\vec{w},b,\vec{\xi}} C\sum_{i=1}^{n} \xi_i + \frac{1}{2}\|\vec{w}\|^2$

with $y_i(\langle \vec{w}, \vec{x}_i \rangle + b) \ge 1 - \xi_i$ and $\xi_i \ge 0$

Dual form: $\max_{\vec{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \langle \vec{x}_i, \vec{x}_j \rangle$

with $0 \le \alpha_i \le C$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$

Maximum Margin Hyperplane:

$f(\vec{x}) = \sum_{i=1}^{n} y_i \alpha_i \langle \vec{x}_i, \vec{x} \rangle + b$ with $b = -\frac{1}{2}\left(\max_{j: y_j=-1}\left(\sum_{i=1}^{n} y_i\alpha_i \langle \vec{x}_i, \vec{x}_j \rangle\right) + \min_{j: y_j=1}\left(\sum_{i=1}^{n} y_i\alpha_i \langle \vec{x}_i, \vec{x}_j \rangle\right)\right)$

$\vec{w} = \sum_{i=1}^{n} y_i \alpha_i \vec{x}_i$

Artificial Neural Networks

Perceptron rule:

if $y_i(\vec{w}_k^T\vec{x}_i) \le 0$: $\vec{w}_{k+1} = \vec{w}_k + y_i\vec{x}_i$

Delta rule: $\vec{w} = \vec{w} + \Delta\vec{w}$, $\Delta\vec{w} = -\eta\nabla E[\vec{w}]$

Linear: $\frac{\partial E}{\partial w_i} = -\sum_{(x,y)\in D}(y - \vec{w}^T\vec{x})x_i$

Sigmoid: $\frac{\partial E}{\partial w_i} = -\sum_{(x,y)\in D}(y - o(\vec{x}))o(\vec{x})(1 - o(\vec{x}))x_i$

Theorem: Block-Novikoff

If: 1. $\forall i \in [n]: \|\vec{x}_i\| \le R$ for some $R \in \mathbb{R}$

2. $\exists \vec{u} \in \mathbb{R}^d, \gamma > 0: \forall i \in [n]: y_i(\vec{u}^T x_i) \ge \gamma$

Then: #updates $k \le \left(\frac{R}{\gamma}\right)^2$

Bayesian Learning

MAP: $h_{MAP} = \underset{h\in\mathcal{H}}{\mathrm{argmax}}\ P(h|D)$

$= \underset{h\in\mathcal{H}}{\mathrm{argmax}} \frac{P(D|h)P(h)}{P(D)}$

$= \underset{h\in\mathcal{H}}{\mathrm{argmax}}\ P(D|h)P(h)$

Likelihood. $L(h) = P(D|h) = \prod_{(x,y)\in D} P(y|x,h)P(x)$

Max. Likelihood h: $h_{ML} = \underset{h\in\mathcal{H}}{\mathrm{argmax}} \sum_{(x,y)\in D} \log P(y|x,h)$

Bayes opt. class.: $y_{Bay.opt.} = \underset{y\in Y}{\mathrm{argmax}} \sum_{h\in\mathcal{H}} P(y|h)P(h|x)$

Naive Bayes class.: $y_{NB} = \underset{y\in Y}{\mathrm{argmax}}\ P(y)\prod_{i=1}^{n} P(a_i|y)$

m-estimate smoothing: $\hat{P}(a_i|y) = \frac{n_c + mp}{n + m}$

Laplace: $m = |A_i|$ $p = \frac{1}{m}$

Logistic Regression: ($\Leftrightarrow$ Linear Regression)

logit: $\log\frac{p}{1-p}$ prob: $\frac{1}{1+e^{-z}}$

Hypothesis: $\sigma(\Theta^T x')$ $\sigma = \frac{1}{1+e^{-z}}$

Determine Θ: Minimize:

$L(\Theta) = -\sum_{(x,y)\in D} y \cdot \log h_\Theta(x) + (1-y)\log(1 - h_\Theta(x))$

Partial derivatives: $\sum_{(x,y)\in D}\left(\frac{1}{1+\exp(-\Theta^T x')} - y\right)x_i$

Comparing Learning Alg.

True error: $\Pr_{x\in D}[f(x) \ne h(x)]$

Sample error: $\frac{1}{|S|}\sum_{x\in S}\delta(f(x), h(x))$ $\delta = 1$, if $f(x) \ne h(x)$

Estimator: Y estimates p of unknown dis.

Bias: $\mathbb{E}[Y] - p$ Unbiased: $\mathrm{Bias}(Y) = 0$

MSE: $\mathbb{E}[(Y-p)^2]$

Variance: $\mathbb{E}[(Y - \mathbb{E}[Y])^2]$

Theorem: $MSE(Y) = Var(Y) + Bias(Y)^2$

Central Limit Theorem:

Sample mean: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

For $n \to \infty$: $\bar{Y} \to \mu$ $\sigma \to \frac{\sigma}{n}$

True error estimator:

Estimator: Sample error

$error_{S_i}(h) = \frac{r}{n}$

Binom. Distribution:

$P(r) = \binom{n}{r} error_D(h)^r (1 - error_D(h))^{n-r}$

Mean: $\mathbb{E}[X] = \sum_{i=0}^{n} i P(i) = np$

Var.: $Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = np(1-p)$

Std.: $\sigma_X = \sqrt{np(1-p)}$

Theorem: $\mathbb{E}[error_S(h)] - error_D(h) = 0$

De Moivre-Laplace Theorem:

$\mu = np$ $\sigma = \sqrt{np(1-p)}$

Z-Scores: $z = \frac{error_S(h) - \mu_{error_S(h)}}{\sigma_{error_S(h)}} \sim \mathcal{N}(0,1)$

%	50	68	80	90	95	98	99
z	.67	1.	1.28	1.64	1.96	2.33	2.58

68-95-99.7 rule:
68.27%: $[\mu - \sigma, \mu + \sigma]$
95.45%: $[\mu - 2\sigma, \mu + 2\sigma]$
99.73%: $[\mu - 3\sigma, \mu + 3\sigma]$

Confidence interval for true error:

$error_D(h) \in [error_S(h) - z\cdot\sigma, error_S(h) + z\cdot\sigma]$

Difference between hypothesis:

Estimator: $\hat{d} = error_{S_1}(h_1) - error_{S_2}(h_2)$

Approx.: $\mu = d$,

$\sigma^2 \approx \frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}$

Confidence interval: $\hat{d} \pm z_N\sqrt{\sigma^2}$

Comparing Learning Algorithms:

Estimate: $\mathbb{E}_{S\subset D}[error_D(L_A(S)) - error_D(L_B(S))]$

Paired t-Test:

Estimate: $\bar{\delta} = \mathbb{E}_{S'\subset S}[error_D(L_A(S')) - error_D(L_B(S'))]$

for S' with $|S'| = \frac{k-1}{k}|S|$

Confidence interval: $\bar{\delta} \pm t_{N,k-1} s_{\bar{\delta}}$

with: $s_{\bar{\delta}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$

$t_{N,k-1}$: t-distribution

Concept Learning:

Confusion matrix:

		predicted	
		+	−
actual	+	TP	FN
	−	FP	TN

Accuracy: $\frac{TP + TN}{TP + FN + FP + TN}$

Precision: $\frac{TP}{TP + FP}$

Recall: $\frac{TP}{TP + FN}$

F-score: $2 \cdot \frac{precision \cdot recall}{precision + recall}$

F-Beta score: $(1+\beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$

Hidden Markov Models

States: $S = \{S_1, S_2, \dots, S_N\}$

State at time t: $q_t \in S$

State Change: $Pr(q_{t+1} = S_j \mid q_t = S_i, \dots, q_1 = S_L)$

Transition Probability Matrix A ($N \times N$)

$a_{ij} = Pr(q_{t+1} = S_j \mid q_t = S_i) \geq 0$, s.t. $\sum_{j=1}^{N} a_{ij} = 1$

Initial Probabilities Π: $\Pi = (\pi_1, \dots, \pi_N)^T$

with $\pi_i = Pr(q_1 = S_i)$

Emission Probability B ($N \times M$):

$b_j(m) = Pr(O_t = V_m \mid q_t = S_j) \geq 0$

with: $\sum_{m=1}^{M} b_j(m) = 1$

Markov Assumption:

$Pr(q_{t+1} = S_j \mid q_t = S_i, \dots, q_1 = S_L) = Pr(q_{t+1} = S_j \mid q_t = S_i)$