1 Basics

$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad \mathcal{N}(x|\mu,\sigma)$

$f(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \quad \mathcal{N}(x|\mu,\Sigma)$

Condition number: $\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$

f(x) on a: $f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots$

Binomial: $f(k,n,p) = Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

$\ln(p(x|\mu,\Sigma)) = -\frac{d}{2}\ln(2\pi) - \frac{\ln|\Sigma|}{2} - \frac{1}{2}(x-\mu)^T\Sigma(x-\mu)$

$X \sim \mathcal{N}(\mu,\Sigma),\ Y = A + BX \Rightarrow Y \sim \mathcal{N}(A + B\mu, B\Sigma B^T)$

// General p-norm: $||x||_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$

Moments

- $Var[X] = \int_x (x-\mu)^2 p(x) dx$
- $Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$
- $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$
- $Cov[X,Y] = E[(X - E[X])(Y - E[Y])]$
- $Cov[aX, bY] = abCov[X,Y]$
- $K_{XY} = cov(X,Y) = E[XY^T] - E[X]E[Y^T]$

Calculus

- Part.: $\int u(x)v'(x)dx = u(x)v(x) - \int v(x)u'(x)dx$
- Chain r.: $\frac{f(y)}{g(x)} = \frac{dz}{dx}\Big|_{x=x_0} = \frac{dz}{dy}\Big|_{z=g(x_0)} \cdot \frac{dy}{dx}\Big|_{x=x_0}$
- $\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{b}) = \mathbf{b}$ • $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x}$
- $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = (\mathbf{A}^\top + \mathbf{A})\mathbf{x} \overset{A\text{ sym.}}{=} 2\mathbf{A}\mathbf{x}$
- $\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top \mathbf{A}\mathbf{x}) = \mathbf{A}^\top \mathbf{b}$ • $\frac{\partial}{\partial \mathbf{X}}(\mathbf{c}^\top \mathbf{X}\mathbf{b}) = \mathbf{c}\mathbf{b}^\top$
- $\frac{\partial}{\partial \mathbf{X}}(\mathbf{c}^\top \mathbf{X}^\top \mathbf{b}) = \mathbf{b}\mathbf{c}^\top$ • $\frac{\partial}{\partial \mathbf{x}}(||\mathbf{x}-\mathbf{b}||_2) = \frac{\mathbf{x}-\mathbf{b}}{||\mathbf{x}-\mathbf{b}||_2}$
- $\frac{\partial}{\partial \mathbf{x}}(||\mathbf{x}||_2^2) = \frac{\partial}{\partial \mathbf{x}}(||\mathbf{x}^\top \mathbf{x}||_2) = 2\mathbf{x}$ • $\frac{\partial}{\partial \mathbf{X}}(||\mathbf{X}||_F^2) = 2\mathbf{X}$
- $x^T A x = Tr(x^T A x) = Tr(x x^T A) = Tr(A x x^T)$
- $\frac{\partial}{\partial A} Tr(AB) = B^T$ • $\frac{\partial}{\partial A} log|A| = A^{-T}$
- $\text{sigmoid}(x) = \sigma(x) = \frac{1}{1+\exp(-x)}$
- $\nabla\text{sigmoid}(x) = \text{sigmoid}(x)(1 - \text{sigmoid}(x))$
- $\nabla\tanh(x) = 1 - \tanh^2(x)$ • $tanh x = \frac{sinh x}{cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

Probability / Statistics

Bayes' Rule $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad \frac{P(X|Y)P(Y)}{\sum_{i=1}^{k} P(X|Y_i)P(Y_i)}$

MGF $\mathbf{M}_X(t) = \mathbb{E}[e^{\mathbf{t}^T\mathbf{X}}]$, $\mathbf{X} = (X_1, .., X_n)$

Jensen's inequality

X:random variable & φ:convex function $\rightarrow$ $\varphi(\mathbb{E}[X]) \le \mathbb{E}[\varphi(X)]$

2 Regression

Estimation

Consistency: $\hat{\theta_n} \overset{P}{\rightarrow} \theta$, i.e. $\forall \epsilon P\{|\hat{\theta_n} - \theta| \ge \epsilon\} \overset{n\rightarrow\infty}{\longrightarrow} 0$

Asymptotic normality: $\sqrt{N}(\theta - \hat{\theta_n}) \rightarrow \mathcal{N}(0, J^{-1}IJ^{-1})$

Asymptotic efficiency: $\hat{\theta_n}$ has the smallest variance among all possible consistent estimators (for large enough N), i.e. $\lim_{n\rightarrow\infty}(V[\hat{\theta_n}]I(\theta))^{-1} = 1$ $\hat{\theta}_{MAP} := \arg\max_\theta \left\{\sum_{i=1}^n log(p(x_i|\theta) + log(p(\theta))\right\}$

Rao-Cramer

$\Lambda = \frac{\partial \log \mathbb{P}(x|\theta)}{\partial\theta}$ (score function), $E[\Lambda] = 0$

Fisher information: $I = \mathbb{V}[\Lambda]$

$J = E[\Lambda^2] = -E[\frac{\partial^2 \log \mathbb{P}(x|\theta)}{\partial\theta\partial\theta^T}] = -E[\frac{\partial\Lambda}{\partial\theta}]$

variance of an estimator is bounded from below by the inverse of Fisher information

MSE bound: $E[(\hat{\theta} - \theta)^2] \ge \frac{[1+b'(\theta)]^2}{nE[\Lambda^2]} + b_{\hat{\theta}}^2$

Biased estimators: $var(\hat{\theta}) \ge \frac{[1+b'(\theta)]^2}{I(\theta)}$

Efficiency: $e(\hat{\theta}) = \frac{I(\theta)^{-1}}{var(\hat{\theta})} \le 1$

Cauchy-Schwarz: $|E(X,Y)|^2 \le E(X^2)E(Y^2)$

Regularized regression

Error: $\hat{R}(w) = \sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda||w||_2^2$ (Ridge)

Closed form: $w^* = (X^T X + \lambda I)^{-1} X^T y$ (Ridge)

Shrinkage: $Xw^* = \sum_{j=1}^d u_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j^T y$, $X = U\Sigma V^T$

LASSO: $w^* = \underset{w}{\text{argmin}} \sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda||w||_1$

Bayesian linear regression

Model: $y = X^T\beta + \epsilon$, with $\epsilon \sim \mathcal{N}(\epsilon|0,\sigma^2 I)$ or $P(y \mid X, \beta, \sigma) = \mathcal{N}(y|X^T\beta, \sigma^2 I)$ $P(\beta|\Lambda) = \mathcal{N}(\beta|0, \Lambda^{-1})$, Post: $P(\beta|X,y,\Lambda) = \mathcal{N}(\beta|\mu_\beta, \Sigma_\beta)$ $\mu_\beta = (X^T X + \sigma^2\Lambda)^{-1} X^T y$, $\Sigma_\beta = \sigma^2(X^T X + \sigma^2\Lambda)^{-1}$ Prediction: $y_{new} = \hat{\beta}_{MAP4pt}^T x_{new} = \mu_\beta^T x_{new}$ $P(y_{new}|x_{new}, X, y, \beta) = \mathcal{N}(\mu_\beta^T x_{new}, \sigma^2 + x_{new}^T \Sigma_\beta x_{new})$

Combination of Regression Models:

$\text{bias}[\hat{f}(x)] = \frac{1}{B}\sum_{i=1}^B \text{bias}[\hat{f}_i(x)]$

$\text{Var}[\hat{f}(x)] = \frac{1}{B^2}\sum_i \text{Var}[\hat{f}_i(x)] + \frac{1}{B^2}\sum_{i,j:i\neq j}$ $\text{Cov}[\hat{f}_i(x), \hat{f}_j(x)] \approx \frac{\sigma^2}{B}$

RSS Estimator

$\hat{\beta} \sim \mathcal{N}(\beta, (X^T X)^{-1}\sigma^2)$.

Bias vs. Variance

$\mathbb{E}_D\mathbb{E}_{X,Y}\left(\hat{f}(X) - Y\right)^2 =$

$\mathbb{E}_D\mathbb{E}_X\left(\hat{f}(X) - \mathbb{E}(Y|X)\right)^2 + \mathbb{E}_{X,Y}(Y - \mathbb{E}(Y|X))^2$

$= \mathbb{E}_X\mathbb{E}_D\left(\hat{f}(X) - \mathbb{E}_D(\hat{f}(X))\right)^2$ (variance)

$+ \mathbb{E}_X\left(\mathbb{E}_D(\hat{f}(X)) - \mathbb{E}(Y|X)\right)^2$ (bias^2)

$+ \mathbb{E}_{X,Y}(Y - \mathbb{E}(Y|X))^2$ (noise)

Ridge Parametric to nonparametric

Ansatz: $w = \sum_i \alpha_i x$

$w^* = \underset{w}{\text{argmin}} \sum_i (w^T x_i - y_i)^2 + \lambda||w||_2^2 =$ $\text{argmin}_{\alpha_{1:n}} \sum_{i=1}^n (\sum_{j=1}^n \alpha_j x_j^T x_i - y_i)^2 + \lambda\sum_i\sum_j \alpha_i\alpha_j(x_i^T x_j)$

$= \text{argmin}_{\alpha_{1:n}} \sum_{i=1}^n (\alpha^T K_i - y_i)^2 + \lambda\alpha^T K\alpha$

$= \text{argmin}_\alpha ||\alpha^T K - y||_2^2 + \lambda\alpha^T K\alpha$

Closed form: $\alpha^* = (K + \lambda I)^{-1} y$

Prediction: $y^* = w^{*T} x = \sum_{i=1}^n \alpha_i^* k(x_i, x)$

3 Gaussian Processes

Gaussian Process

$[y_1, y_2, \ldots]^T = X\beta + \epsilon \sim \mathcal{N}(y|0, X\Lambda^{-1}X^T + \sigma^2 I)$

$y \sim \mathcal{N}(y|m(X), K(X,X) + \sigma^2 I) = P(y|X,\Theta)$

$\begin{bmatrix} y \\ y_{n+1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} y \\ y_{n+1} \end{bmatrix} \middle| \begin{bmatrix} m(X) \\ m(x_{n+1}) \end{bmatrix}, \begin{bmatrix} C_n & k \\ k^T & c \end{bmatrix}\right)$

$p(y_{n+1}|x_{n+1}, X, y)) = \mathcal{N}(y_{n+1}|\mu_{n+1}, \sigma_{n+1}^2)$

$\mu_{n+1} = m(x_{n+1}) + k^T C_n^{-1}(y - m(X))$

$\sigma_{n+1}^2 = c - k^T C_n^{-1} k, k = k(x_{n+1}, X)$

$c = k(x_{n+1}, x_{n+1}) + \sigma^2, C_n = K_n + \sigma^2 I$

GP Hyperparameter Optimization

Log-likelihood:

$l(Y|\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|C_n| - \frac{1}{2}Y^T C_n^{-1} Y$

Set of hyperparameters θ determine parameters C_n. Gradient descent: $\nabla_{\theta_i} l(Y|\theta) = -\frac{1}{2}tr(C_n^{-1}\frac{\partial C_n}{\partial\theta_i}) + \frac{1}{2}Y^T C_n^{-1}\frac{\partial C_n}{\partial\theta_i}C_n^{-1}Y$

Kernels

$K(x,y) = <\phi(x), \phi(y)>$ for some feature mapping $\phi(x)$

Psd Gram Matrix: $c^T K c \ge 0, \sum_i\sum_j c_i c_j k(x_i, x_j) \ge 0$

All principal minors of K need $det \ge 0$;

$k(x,y) = k(y,x)$; $k(x,x) \ge 0$; $k(x,x)k(v,v) \ge k(x,y)^2$ Closure Properties: psd prop. closed under pointwise limits (since each K_n is a kernel)

$k(x,y) = k_1(x,y) + k_2(x,y)$, $k(x,y) = k_1(x,y)k_2(x,y)$

$k(x,y) = f(x)f(y)$, $k(x,y) = k_3(\phi(x), \phi(y))$

$k(x,y) = \exp(\alpha k_1(x,y)), \alpha > 0, |X \cap Y| = kernel$

$k(x,y) = p(k_1(x,y))$, $p(\cdot)$ polynomial with pos. coeff.

$k(x,y) = k_1(x,y)/\sqrt{(k_1(x,x)k_1(y,y)}$

Gaussian (rbf): $k(x,y) = \exp(-\frac{||x-y||^2}{2\sigma^2})$ inf.dim.

Sigmoid: $k(x,y) = \tanh(k \cdot x^T y - b)$ not valid for $\forall k, b$

Polynomial: $k(x,y) = (x^T y + c)^d, d \in N, c \ge 0$

Periodic: $k(x,y) = \sigma^2 exp(\frac{2\sin^2(\pi|x-y|/p)}{\ell^2})$

4 Numerical Estimating Methods

Actual Risk: $\mathcal{R}(f) := \mathbb{E}_{x,y}[(y - f(x))^2]$

Empiricial Risk: $\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_i (y_i - f(x_i))^2$

Generalization Error: $G(f) = |\hat{\mathcal{R}}(f) - \mathcal{R}(f)|$

K-fold cross validation

$\hat{f}^{-\nu} \in \arg\min_f \frac{1}{|Z^{-\nu}|}\sum_{i \in Z^{-\nu}}(y_i - f(x_i))^2$

$\hat{\mathcal{R}}^{cv} = \frac{1}{n}\sum_i (y_i - \hat{f}^{-\kappa(i)}(x_i))^2$, $k(i)$ is fold i^{th} fold

Problem: systematic tendency to underfit.

Leave-one-out

unbiased, high variance

$\hat{f}^{-i} \in \arg\min_f \frac{1}{n-1}\sum_{j:j\neq i} L(y_i, f(x_i))$

$\hat{\mathcal{R}}^{LOOCV} = \frac{1}{n}\sum_i L(y_i, \hat{f}^{-i}(x_i))$

Bootstrapping

Resampling with replacement from data D to produce B boostrap datasets D^{*b}. $S(D)$ is expected generalization error of prediction model trained on D. Var: $\sigma^2(S) = \frac{1}{B-1}\sum_{b=1}^B (S(D^{*b}) - \overline{S})^2$ with mean: $\hat{R}_{boot}(f) = \overline{S} = \frac{1}{B}\sum_{b=1}^B(\frac{1}{N}\sum_{i=1}^N L(y_i, \hat{f}_{D^{*b}}(x_i)))$ with $\hat{f}_{D^{*b}}(x_i)$ being the prediction model. $\hat{R}_{boot}^{LOO}(f) = \frac{1}{N}\sum_{i=1}^N \frac{1}{|C^{-i}|}\sum_{b \in C^{-i}} L(y_i, \hat{f}_{D^{*b}}(x_i))$ where C^{-i} denotes the set of bootstrap sets not containing data point i. Note: L can be $I_{\{c(x_i)\neq y_i\}}$. $\hat{R}_{boot}$ is optimistic. Hence use: $\hat{R}^{.0632} = 0.368\hat{R}_{boot} + 0.632\hat{R}_{boot}^{(LOO)}$.

Prob. not to appear in set: $(1 - \frac{1}{n})^n = \frac{1}{e}$ for $n \rightarrow \infty$

Jackknife

Goal: Numerical estimate of bias of an estimator $\hat{S}_n$. Jackknife estimator: $\hat{S}^{JK} = \hat{S}_n - bias^{JK}$ with $bias^{JK} = (n-1)(\tilde{S}_n - \hat{S}_n)$ with $\tilde{S}_n = \frac{1}{n}\sum_{i=1}^n \hat{S}_{n-1}^{(-i)}$ with $\hat{S}_{n-1}^{(-i)}$ being the leave-1-out estimator.

Information Criteria

$BIC = ln(n)k - 2ln(\hat{L})$, $AIC = 2k - 2ln(\hat{L})$

$TIC = 2trace[I_1(\theta_k)J_1^{-1}(\theta_k)] - 2ln(\hat{L})$, where k: num. params, n: num. data points, likelihood: $\hat{L} = p(X|\theta_k, M)$

5 Classification

Loss-Functions

True class: $y \in \{-1, 1\}$, pred. $z \in [-1, 1]$

Cross-entropy (log loss): ($y' = \frac{(1+y)}{2}$ and $z' = \frac{(1+z)}{2}$) $L(y', z') = -[y'log(z') + (1-y')log(1-z')]$

Hinge Loss: $L(y,z) = max(0, 1 - yz)$

Perceptron Loss: $L(y,z) = max(0, -yz)$

Logistic loss: $L(y,z)=log(1+exp(-yz))$
Square loss: $L(y,z)=\frac{1}{2}(1-yz)^2$
Exponential loss: $L(y,z)=exp(-yz)$
Binomial deviance: $L(y,z)=1+exp(-2yz)$
0/1 Loss: $L(y,z)=\mathbb{I}\{sign(z)\neq y\}$

Perceptron
Gradient descent: $a(k+1)=a(k)-\eta(k)\nabla J(a(k))$
$J(a)\approx J(a(k))+\nabla J^T(a-a(k))+\frac{1}{2}(a-a(k))^T H(a-a(k))$, $H=\frac{\partial^2 J}{\partial a_i \partial a_j}$
2^{nd} order algorithm: $\eta_{opt}=\frac{\|\nabla J\|^2}{\nabla J^T H \nabla J}$
Newton's rule: $a(k+1)=a(k)-H^{-1}\nabla J$
Perceptron criteria: $J_p(a)=\sum_{\tilde{x}\in\tilde{\mathcal{X}}^{mc}}(-a^T\tilde{x})$
Perceptron rule: $a(k+1)=a(k)+\eta(k)\sum_{\tilde{x}\in\tilde{\mathcal{X}}^{mc}}\tilde{x}$
Perceptron convergence: $\|a(k+1)-\alpha\hat{a}\|^2 = \|a(k)-\alpha\hat{a}\|^2 + 2(a(k)-\alpha\hat{a})^T\tilde{x}^k + \|\tilde{x}^k\|^2 \le \|a(k)-\alpha\hat{a}\|^2 - 2\alpha\gamma + \beta^2$ where $\beta^2 = max_i\|\tilde{x}_{i\in\tilde{\mathcal{X}}^{mc}}\|^2$ and $\gamma = min_{i\in\tilde{\mathcal{X}}^{mc}}(\hat{a}^T\tilde{x}_i)>0$ for $\alpha=\beta^2/\gamma$ then $k_0=\alpha^2\|\hat{a}\|^2/\beta^2=\beta^2\|\hat{a}\|^2/\gamma^2$

6 Design of Discriminant
Fisher's Linear Discriminant:
$\mathbb{R}^d\to\mathbb{R}^{(k-1)}$: $\vec{y}_i=\vec{w}_i^T\vec{x}, 1\le i\le k-1, \vec{y}=W^T\vec{x}$
Criterion: $J(W)=\frac{|W^T\Sigma_B W|}{|W^T\Sigma_W W|}\overset{2\text{ classes}}{=}\frac{(m_2-m_1)^2}{s_1^2+s_2^2}\xrightarrow[d/dW=0]{\text{maximize}}$
$\Sigma_B=\sum_i n_i(m_i-m)(m_i-m)^T$ (Between class variance)
$\Sigma_W=\sum_i\sum_{x\in X_i}(x-m_i)(x-m_i)^T$ (Within class variance)
$m_i=\frac{1}{n_i}\sum_{x\in X_i}x$, $m=\frac{1}{n}\sum_x x$
solution: $\hat{w}\overset{2\text{ classes}}{=}\Sigma_W^{-1}(m_1-m_2)$

7 SVM
Primal Problem: ($C\to\infty$: Hard Margin)
$\min_w \frac{1}{2}w^Tw+C\sum_{i=1}^n\xi_i$ s.t. $z_i(w^T\phi(y_i)+w_0)\ge 1-\xi_i, \xi_i\ge 0$
Dual Problem: : $L(w,w_0,\xi,\alpha,\beta)=\frac{1}{2}w^Tw+C\sum_{i=1}^n\xi_i-\sum_{i=1}^n\beta_i\xi_i-\sum_{i=1}^n\alpha_i(z_i(w^T\phi(y_i)+w_0)-1+\xi_i)$
$\max_\alpha L(a)=\sum_{i=1}^n\alpha_i-\frac{1}{2}\sum_{i,j=1}^n z_iz_j\alpha_i\alpha_j\phi(y_i,y_j)$
s.t. $\sum_{j=1}^n z_j\alpha_j=0\wedge C\ge\alpha_i\ge 0$
optimal hyperplane: $w^*=\sum_{i=1}^n\alpha_i^*z_i\phi(y_i)$
$w_0^*=\frac{1}{n_s}\sum_{i\in S}(z_n-\sum_{j\in S}\alpha_jz_j\phi(y_i,y_j))$
$\overset{\text{linear}}{=}-\frac{1}{2}(min_{i:z_i=1}w^{*T}y_i+max_{i:z_i=-1}w^{*T}y_i)$
Only for support vectors: $\alpha_i^*>0$
Prediction: $z(y)=sign(\sum_{i=1}^n\alpha_iz_i\phi(y,y_i)+w_0)$
$\overset{\text{linear}}{=}sign(w^{*T}x+w_0)$
Homog. Coordinates: condition $\sum_{j=1}^n z_j\alpha_j=0$ falls away.

8 Non-linear SVM
Multiclass SVM
$\min_{w,\eta\ge 0}\frac{1}{2}w^Tw+C\sum_i\xi_i$
s.t. $\forall y_i\in Y:(w_{z_i}^Ty_i)-max_{z\neq z_i}(w_z^Ty_i)\ge 1-\xi_i$
Structured SVM
$\min_{w,\eta}\frac{1}{2}\|w\|^2+\frac{1}{n}\sum_{i=1}^n\eta_i$, $\eta\ge H_i(w)\forall i$, where $H_i(w)=\max_{y\in Y(x_i)}L(y_i,y)-w^T(\phi(x_i,y_i)-\phi(x_i,y))$

9 Ensemble method
Random Forest
for b=1:B do:
draw a bootstrap sample D_b
repeat until node size<n_{min}:
1. select m features from p features
2. pick the best variable and split-point
3. Split the node accordingly

return the forest $\{\hat{c}_b(x)\}_{b=1}^B$
Boosting: Train weak learners sequentially on all data, but reweight misclassifed samples higher, Bias ↓
Adaboost
Initialize weights $w_i=1/n$, for b=1:B do:
1. Fit classifier $c_b(x)$ with weights w_i
2. Compute error $\epsilon_b=\sum_i w_i^{(b)}\mathbb{1}_{[c_b(x_i)\neq y_i]}/\sum_i w_i^{(b)}$
3. Compute coeff. $\alpha_b=log(\frac{1-\epsilon_b}{\epsilon_b})$
4. Update weights $w_i=w_i\exp(\alpha_b\mathbb{1}_{[y_i\neq c_b(x_i)]})$

Return $\hat{c}_B(x)=\text{sign}\left(\sum_{b=1}^B\alpha_bc_b(x)\right)$
Loss: Exponential loss function
Model: Additive logistic regression
Bayesian approach (assumes posteriors)
Newtonlike updates (Gradient Descent)
Bagging
return ensemble class. $\hat{c}_B(x)=sgn(\sum_{i=1}^B c_i(x))$
Works: Covariance small (different subset for training), Variance small (similar behaviour of weak learners), biases weakly affected.
Bias↓&Var.↓: Use complex decision tree (bias↓), ensemble mult. decision trees (var↓)
Gaussian Mixtures
Estimate $\hat{\theta}=\{\mu_1,...,\mu_k,\Sigma_1,...,\Sigma_k\}$ that maximize the likelihood of sample feature vectors $\mathcal{X}=\{x_1,...,x_n\}$:
$p(\mathcal{X}|\pi_i,...,\pi_k,\theta_1,...,\theta_k)=\prod_{x\in\mathcal{X}}\sum_{c\le k}\pi_cp(x|\theta_c)$
Log-Likelihood: $L(\mathcal{X}|\pi,\theta)=\sum_{x\in\mathcal{X}}\log\sum_{c\le k}\pi_cp(x|\theta_c)$
Expectation Maximization
$L(\mathcal{X},M|\theta)=\sum_{x\in\mathcal{X}}\sum_{c=1}^k M_{xc}\log(\pi_cP(x|\theta_c)$
$Q(\theta;\theta^{(j)})=\mathbb{E}_M[L(\mathcal{X},M|\theta)|\mathcal{X},\theta^{(j)}]$, M latent variable
$M_{xc}=1$ if cluster c has generated x, else $M_{xc}=0$
$\mathbb{E}_M[M_{xc}|\mathcal{X},\theta^{(j)}]=P(M_{xc}=1)=P(c|x,\theta^{(j)})=\frac{P(x|c,\theta^{(j)})P(c|\theta^{(j)})}{P(x|\theta^{(j)})}=\frac{\pi_cP(x|c,\theta^{(j)})}{\sum_{c=1}^K\pi_cP(x|c,\theta^{(j)})}=:\gamma_{xc}$

while not converged **do**
  E-Step: Compute γ_{xc} for all x,c Compute $m_c:=\sum_x\gamma_{xc}$ for all c
  M-Step: max $Q(\theta;\theta^{(j)})$ s.t. $\sum_c\pi_c=1$
  $\mu_c^{(j+1)}=\frac{\sum_{x\in\mathcal{X}}\gamma_{xc}x}{m_c}$ $\pi_c^{(j+1)}=\frac{1}{|\mathcal{X}|}m_c$
  $\Sigma_c^{(j+1)}=\frac{\sum_{x\in\mathcal{X}}\gamma_{xc}(x-\mu_c)(x-\mu_c)^T}{m_c}$
end while

Lagrangian with fixed γ_{xc}
$L=\sum_x\sum_c\gamma_{xc}\log(\pi_cP(x|c,\theta_c))-\lambda(\sum_c\pi_c-1)$
For GMM: $P(x|c,\theta^{(j)})=\mathcal{N}(x|\mu_c,\Sigma_c)$

10 Neural Network
Backpropagation
For each unit j on the output layer:
- Compute error signal: $\delta_j=\ell_j'(f_j)$
- For each unit i on layer L: $\frac{\partial}{\partial w_{j,i}}=\delta_jv_i$

For each unit j on hidden layer $l=\{L-1,..,1\}$:
- Error signal: $\delta_j=\phi'(z_j)\sum_{i\in Layer_{l+1}}w_{i,j}\delta_i$
- For each unit i on layer $l-1$: $\frac{\partial}{\partial w_{j,i}}=\delta_jv_i$

11 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.: $\mathcal{R}(\hat{c}_n^*)-\inf_{c\in\mathcal{C}}\mathcal{R}(c)\le 2\sup_{c\in\mathcal{C}}|\hat{\mathcal{R}}_n(c)-\mathcal{R}(c)|$
$P\{\mathcal{R}(\hat{c}_n^*)-\mathcal{R}(c^*)>\epsilon\}\le P\{\sup_{c\in\mathcal{C}}|\hat{\mathcal{R}}_n(c)-\mathcal{R}(c)|>\frac{\epsilon}{2}\}$
$\le 2|\mathcal{C}|exp(-2n\epsilon^2/4)\le 8s(\mathcal{A},n)exp(-n\epsilon^2/32)$ and $s(\mathcal{A},n)\le n^{\mathcal{V}_{\mathcal{A}}}$
Markov ineq: $P\{X\ge\epsilon\}\le\frac{\mathbb{E}[X]}{\epsilon}$ (for nonneg. X)
Boole's inequality: $P(\bigcup_iA_i)\le\sum_iP(A_i)$
Hoeffding's lemma: $\mathbb{E}[e^{sX}]\le exp(\frac{1}{8}s^2(b-a)^2)$ where $\mathbb{E}[X]=0$, $P(X\in[a,b])=1$
Hoeffding's: $P\{S_n-\mathbb{E}[S_n]\ge t\}\le exp(-\frac{2t^2}{\sum_i(b_i-a_i)^2})$
Normalized: $P\{\tilde{S}_n-\mathbb{E}[\tilde{S}_n]\ge\epsilon\}\le exp(-\frac{2n^2\epsilon^2}{\sum_i(b_i-a_i)^2})$
Error bound: $P\{\sup_{c\in\mathcal{C}}|\hat{\mathcal{R}}_n(c)-\mathcal{R}(c)|>\epsilon\}\le 2|\mathcal{C}|exp(-2n\epsilon^2)$

The $\mathcal{VC}$ dimension of a model f is the maximum number of points that can be arranged so that f shatters them.

12 Nonparametric Bayesian methods
$Dir(x|\alpha)=\frac{1}{B(\alpha)}\prod_{k=1}^n x_k^{\alpha_k-1}$, $B(\alpha)=\frac{\prod_{k=1}^n\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^n\alpha_k)}$
$\mathbb{E}[1]=\sum_{i=1}^N\frac{\alpha}{\alpha+i}\sim(\alpha log(N))$
de Finetti: $p(X_1,...,X_n)=\int(\prod_{i=1}^n p(x_i|G))dP(G)$
$p(z_i=k|\mathbf{z}_{-i},\mathbf{x},\alpha,\boldsymbol{\mu})=\begin{cases}\frac{N_{k,-i}}{\alpha+N-1}p(x_i|\mathbf{x}_{-i,k},\boldsymbol{\mu}) & \exists k\\ \frac{\alpha}{\alpha+N-1}p(x_i|\boldsymbol{\mu}) & \text{otherwise}\end{cases}$

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: $\mathcal{N}(\mu_{z_i},\sigma)$

13 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
$f(x)=\log(\frac{P(y=1|x)}{P(y==1|x)}), y=sign(f(x))$

14 Neural Networks
Learning features
Parameterize the feature maps and optimize over the parameters:
$w^*=\underset{w,\Theta}{\text{argmin}}\sum_{i=1}^n l(y_i,\sum_{j=1}^m w_j\Phi(x_i,\Theta_j))$
Reformulating the perceptron
Ansatz: $w=\sum_{j=1}^n\alpha_jy_jx_j$
$\min_{w\in\mathbb{R}^d}\sum_{i=1}^n\max[0,-y_iw^Tx_i]$
$=\min_{\alpha_{1:n}}\sum_{i=1}^n\max[0,-y_i(\sum_{j=1}^n\alpha_jy_jx_j)^Tx_i]$
$=\min_{\alpha_{1:n}}\sum_{i=1}^n\max[0,-\sum_{j=1}^n\alpha_jy_iy_jx_i^Tx_j]$
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 $\hat{y}=sign(\sum_{j=1}^n\alpha_jy_jk(x_j,x_i))$
If $\hat{y}\neq y_i$ set $\alpha_i=\alpha_i+\eta_t$