1 Probability

Cond. Ind. $X \perp Y|Z \implies P(X,Y \mid Z) = P(X \mid Z)\,P(Y \mid Z)$

Cond. Ind. $X \perp Y|Z \implies P(X \mid Y,Z) = P(X \mid Z)$

$\mathbb{E}[X] = \int_{\mathcal{X}} t \cdot f_X(t)\,dt =: \mu_X$

$\mathrm{Var}[X] = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] = \int_{\mathcal{X}} (t-\mathbb{E}[X])^2 f_X(t)\,dt = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$\mathrm{Cov}(X,Y) = \int_{\mathcal{X}}\int_{\mathcal{X}} f(x,y)(x-\mu_x)(y-\mu_x)\,dx\,dy$

"$\mathbf{X}^2 = \mathbf{X}\mathbf{X}^T$" ≥ 0 ((symmetric) positive semidefinite)

$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$\mathrm{Var}[\mathbf{AX}] = \mathbf{A}\,\mathrm{Var}[\mathbf{X}]\,\mathbf{A}^\mathsf{T}$ $\quad \mathrm{Var}[aX+b] = a^2\,\mathrm{Var}[X]$

$\mathrm{Var}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i^2\,\mathrm{Var}[X_i] + 2\sum_{i,j,i<j} a_i a_j\,\mathrm{Cov}(X_i,X_j)$

$\mathrm{Var}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i^2\,\mathrm{Var}[X_i] + \sum_{i,j,i\neq j} a_i a_j\,\mathrm{Cov}(X_i,X_j)$

$\frac{\partial}{\partial t} P(X \leq t) = \frac{\partial}{\partial t} F_X(t) = f_X(t)$ (derivative of c.d.f. is p.d.f)

$f_{\alpha Y}(z) = \frac{1}{\alpha} f_Y(\frac{z}{\alpha})$

Empirical CDF: $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \leq t\}}$

Empirical PDF: $\hat{f}_n(t) = \frac{1}{n}\sum_{i=1}^n \delta(t - X_i)$ (continuous)

Empirical PDF: $\hat{p}_n(t) = \frac{1}{n}|x = t|\, x \in D$ (discrete)

T. The MGF $\psi_X(t) = \mathbb{E}\left[e^{tX}\right]$ characterizes the distr. of a rv

$Be(p)$: $\quad pe^t + (1-p)$	$\mathcal{N}(\mu,\sigma)$: $\quad \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$
$Bin(n,p)$: $(pe^t + (1-p))^n$	$Gam(\alpha,\beta)$: $\left(\frac{1}{a-\beta t}^{\alpha}\right)$ for $t < 1/\beta$
$Pois(\lambda)$: $\quad e^{\lambda(e^t-1)}$	

T. If $X_1,\dots,X_n$ are ind. rvs with MGFs $M_{X_i}(t) = \mathbb{E}\left[e^{tX_i}\right]$, then the MGF of $Y = \sum_{i=1}^n a_i X_i$ is $M_Y(t) = \prod_{i=1}^n M_{X_i}(a_i t)$.

T. Let X, Y be ind., then the p.d.f. of $Z = X + Y$ is the conv. of the p.d.f. of X and Y:

$f_Z(z) = \int_{\mathbb{R}} f_X(t) f_Y(z-t)\,dt = \int_{\mathbb{R}} f_X(z-t) f_Y(t)\,dt$

$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$

$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}-\hat{\boldsymbol{\mu}})(\mathbf{x}-\hat{\boldsymbol{\mu}})^\mathsf{T}$

T. $P\left(\begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix} \middle| \begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right)$

$\mathbf{a}_1, \mathbf{u}_1 \in \mathbb{R}^e$, $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{e\times e}$ p.s.d. $\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{e\times f}$ p.s.d.

$\mathbf{a}_2, \mathbf{u}_2 \in \mathbb{R}^f$, $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{f\times f}$ p.s.d. $\boldsymbol{\Sigma}_{21} \in \mathbb{R}^{f\times e}$ p.s.d.

$P(\mathbf{a}_2 \mid \mathbf{a}_1 = \mathbf{z}) = \mathcal{N}\left(\mathbf{a}_2 \mid \mathbf{u}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}-\mathbf{u}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right)$

T. (Chebyshev) Let X be a rv with $\mathbb{E}_X[=]\mu$ and variance $\mathrm{Var}[X] = \sigma^2 < \infty$. Then for any $\epsilon > 0$, we have $P(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$.

T. (Cramer Rao Bound) If $\hat{\boldsymbol{\theta}}$ is unb., cons. est., then

$MSE(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] \geq \frac{1}{\mathcal{I}_n(\theta)} > 0, \quad$ where

$\mathcal{I}_n(\theta) = -\sum_{i=1}^n \mathbb{E}\left[\frac{\partial^2 \log(P(x_i \mid \boldsymbol{\theta}))}{\partial\theta^2}\right]$. $\quad$ (Fisher Information)

D. (Conditional Expected Risk) Given rv X

$R(f,X) = \int_{\mathcal{Y}} L(Y,f(X))P(Y \mid X)\,dY$

D. (Total Expected Risk) for rvs X, Y

$R(f) = \mathbb{E}_X[R(f,X)] = \int_{\mathcal{X}} R(f,X)P(X)\,dX = \mathbb{E}_{X,Y}[L(Y,f(X))]$.

2 Matrix Calculus

$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla_{\mathbf{x}} f(\mathbf{x}_0)^\mathsf{T}(\mathbf{x}-\mathbf{x}_0)$ (2nd order Taylor Expan.)
$\quad + \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^\mathsf{T}\mathrm{Hess}(f;\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + \mathcal{O}((\mathbf{x}-\mathbf{x}_0)^3)$

$\mathrm{Hess}(f;\mathbf{x}_0) = \frac{\partial^2 f}{\partial\mathbf{x}\partial\mathbf{x}^\mathsf{T}}(\mathbf{x}_0)$

$\frac{\partial}{\partial\mathbf{x}}[\mathbf{Ax}] = \mathbf{A}^\mathsf{T} \quad \frac{\partial}{\partial\mathbf{x}}[\mathbf{A}\mathbf{f}(\mathbf{x})] = \mathbf{A}^\mathsf{T}\frac{\partial}{\partial\mathbf{x}}[\mathbf{f}(\mathbf{x})]$

$\frac{\partial}{\partial\mathbf{x}}[\mathbf{x}^\mathsf{T}\mathbf{Ax}] = 2\mathbf{A}^\mathsf{T}\mathbf{x} \quad \frac{\partial}{\partial\mathbf{x}}\left[\|\mathbf{f}(\mathbf{x})\|_2^2\right] = 2\frac{\partial}{\partial\mathbf{x}}[\mathbf{f}(\mathbf{x})]\,\mathbf{f}(\mathbf{x})$

T. (Sylvester Criterion) A $d \times d$ matrix is positive semi-definite if and only if all the upper left $k \times k$ for $k = 1,\dots,d$ have a positive determinant.

3 Misc

T. (Jensen) f convex/concave, $\forall i$: $\lambda_i \geq 0$, $\sum_{i=1}^n \lambda_i = 1$

$f\left(\sum_{i=1}^n \lambda_i \mathbf{x}_i\right) \leq / \geq \sum_{i=1}^n \lambda_i f(\mathbf{x}_i)$

Special case: $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.

T. (CSU) $|\langle\mathbf{x},\mathbf{y}\rangle|^2 \leq \langle\mathbf{x},\mathbf{x}\rangle \cdot \langle\mathbf{y},\mathbf{y}\rangle$

Special case: $(\sum x_i y_i)^2 \leq (\sum x_i^2)(\sum y_i)^2$

Special case: $\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\,\mathbb{E}[Y^2]$

Lag.: $f(x,y)$ $s.t.$ $g(x,y) = c \quad \mathcal{L}(x,y,\gamma) = f(x,y) - \gamma(g(x,y)-c)$

4 Kernels

D. (Kernel) Kernel functions $k(x,x')$ must satisfy (cf. properties of covariance matrices)

1. Symmetry: $k(x,x') = k(x',x)$
2. Positive semi-definiteness (continuous case):
   $\forall f \in L_2\,\forall\Omega \subset \mathbb{R}$: $\int_\Omega k(x,x')f(x)f(x')\,dx\,dx' \geq 0$
3. Positive semi-definiteness (discrete case):
   For any $n \in \mathbb{N}$, and any set $S = \{\mathbf{x}_1,\dots,\mathbf{x}_n\}$, the kernel (Gram) matrix $\mathbf{K} = (k(\mathbf{x}_i,\mathbf{x}_j))_{i,j=1}^n$ must be positive semidefinite ($\forall\mathbf{x}$: $\mathbf{x}^\mathsf{T}\mathbf{K}\mathbf{x} \geq 0$).

T. (Comp. Rules) $k_1 + k_2$; $k_1 \cdot k_2$; $c \cdot k_1$ $(c > 0)$; $f(k_1)$ where f is a poly. w. pos. coeff., or the exp. func..

Kernel $k(\mathbf{x},\mathbf{y}) =$	**Feature Map** $\boldsymbol{\phi}(\mathbf{x}) =$
$k_1(\mathbf{x},\mathbf{y}) + k_2(\mathbf{x},\mathbf{y})$	$(\boldsymbol{\phi}_1(\mathbf{x}),\boldsymbol{\phi}_2(\mathbf{x}))^\mathsf{T}$
$c \cdot k_1(\mathbf{x},\mathbf{y})$ $(c > 0)$	$\sqrt{c}\cdot\boldsymbol{\phi}_1(\mathbf{x})$
$k_1(\mathbf{x},\mathbf{y}) \cdot k_2(\mathbf{x},\mathbf{y})$	$(\phi_1(\mathbf{x})_i \cdot \phi_2(\mathbf{y})_j)^\mathsf{T}_{\text{for } 1\leq i\leq d_1, 1\leq j\leq d_2}$
$\mathbf{x}^\mathsf{T}\mathbf{Ay}$ $\mathbf{A}$ p.s.d.	$\mathbf{L}^\mathsf{T}\mathbf{x}$, $\mathbf{A} = \mathbf{LL}^\mathsf{T}$
$\mathbf{x}^\mathsf{T}\mathbf{M}^\mathsf{T}\mathbf{Mx}$, $\mathbf{M}$ arbitrary	$\mathbf{Mx}$

Linear $k(x,y) = x^\top y$, Polynomial $k(x,y) = (x^\top y + 1)^d$

RBF $k(x,y) = \exp\left(\frac{-\|x-y\|_2^2}{h^2}\right)$, Sigmoid $k(x,y) = \tanh(kx^\top y - b)$

Proof Tricks: Counterexample, Prove by "for arbitrary n, $x_1,\dots,x_n$, $c_1,\dots,c_n$ $\sum_i\sum_j c_i c_k k(x_i,x_j) \geq 0$", constant kernel $k(x,y) = c$ is a kernel, find feature map and inner product.

5 Regression

Problem What is the optimal estimate of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ based on noisy data $y_i = f(x_i) + \epsilon_i$

Solution the regression function

$$f^*(x) = \mathbb{E}[Y \mid X = x] = \int_{\mathcal{Y}} y \cdot P(y \mid X = x)\,dy$$

5.1 Ridge Regression

$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{y}-\mathbf{Xw}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$ $\quad$ ($\lambda > 0$, chosen via CV)

$\mathbf{w}^* = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$ $\quad$ (always has a solution)

$\mathbf{g}_t = -2\mathbf{X}^\mathsf{T}(\mathbf{y}-\mathbf{Xw}_t) + 2\lambda\mathbf{w}_t$

Note: Scale of the data matters for λ! ($\to$ normalize data)

$Y \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x},\sigma^2)$, $\quad y_i = \mathbf{w}^\mathsf{T}\mathbf{x}_i + \epsilon$, $\quad \epsilon \sim \mathcal{N}(0,\sigma^2)$

$P\left(Y = y \,\middle|\, \mathbf{x} = \mathbf{x}, \boldsymbol{\theta} = (\mathbf{w},\sigma^2)\right) = \mathcal{N}(y; h(\mathbf{x}),\sigma^2)$, $h(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}$

Weights prior: $\mathbf{w} \sim \mathcal{N}(0,\beta^2\mathbf{I})$, $\quad w_i \sim \mathcal{N}(0,\beta^2)$

Maximizing $P(\mathbf{w} \mid D)$ then leads to the connection $\lambda = \frac{\sigma^2}{\beta^2}$.

5.2 Bayesian Linear Regression

$Y = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta} + \epsilon$, $\epsilon \sim \mathcal{N}(0,\sigma^2)$

$P(Y \mid \boldsymbol{X},\boldsymbol{\beta},\sigma) = \mathcal{N}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta},\sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(Y - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})^2\right)$

$P(\boldsymbol{\beta} \mid \boldsymbol{\Lambda}) = \mathcal{N}(\mathbf{o},\boldsymbol{\Lambda}^{-1}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{\beta}\right)$

$P(\boldsymbol{\beta} \mid \mathbf{X},\mathbf{y},\boldsymbol{\Lambda},\sigma) = \frac{P(\mathbf{y} \mid \mathbf{X},\boldsymbol{\beta},\boldsymbol{\Lambda},\sigma)\,P(\boldsymbol{\beta} \mid \mathbf{X},\boldsymbol{\Lambda},\sigma)}{P(\mathbf{y} \mid \mathbf{X},\boldsymbol{\Lambda},\sigma)} \propto P(\mathbf{y} \mid \mathbf{X},\boldsymbol{\beta},\sigma)\,P(\boldsymbol{\beta} \mid \boldsymbol{\Lambda})$

$= \prod_{i=1}^n P(y_i \mid \mathbf{x}_i,\boldsymbol{\beta},\sigma)\,P(\boldsymbol{\beta} \mid \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{\beta};\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$

$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} - \sigma^2\boldsymbol{\Lambda}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \sigma^2\left(\mathbf{X}^\mathsf{T}\mathbf{X} + \sigma^2\boldsymbol{\Lambda}\right)^{-1}$

Special Case: Ridge Regression: $\boldsymbol{\Lambda} = \lambda\mathbf{I}$, $\sigma = 1$

5.3 Kernelized Ridge Regression

Insight optimal $\mathbf{w}^*$ lies in the span of the data.

$\mathbf{w}^* = \mathbf{X}_\phi^\mathsf{T}\mathbf{z}^* \quad (\mathbf{K} = \mathbf{X}_\phi\mathbf{X}_\phi^\mathsf{T} \in \mathbb{R}^{n\times n})$

$\mathbf{z}^* = \arg\min_{\mathbf{z}} \|\mathbf{Kz}-\mathbf{y}\|_2^2 + \lambda\mathbf{z}^\mathsf{T}\mathbf{Kz}$

1) Closed form $\mathbf{z}^* = (\mathbf{X}_\phi\mathbf{X}_\phi^\mathsf{T} + \lambda\mathbf{I})^{-1}\mathbf{y} = (\mathbf{K}-\lambda\mathbf{I})^{-1}\mathbf{y}$

2) Gradient descent $\mathbf{g}_t = 2\mathbf{K}^\mathsf{T}(\mathbf{Kz}-\mathbf{y}) + 2\lambda\mathbf{Kz}$

Prediction $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\phi(\mathbf{x}) = \ldots = \sum_{i=1}^n z_i k(\mathbf{x}_i,\mathbf{x})$

Bayesian Interpretation Same as ridge regression, except that the hypothesis class for $\mathcal{H}$ for h (comp. of mean) may be different.

5.4 Sparse Regression: LASSO

Requires $\|\mathbf{w}\|_1 \leq s$. Prior: $w_i \sim p(w_i;0,b) = \frac{1}{2b}\exp\left(-\frac{|w_i-\mu|}{b}\right)$ where $\mu = 0$, (connection: $\lambda = \frac{2\sigma^2}{b}$).

5.5 Ensemble of Regressions

T. The average of B unbiased estimators $\hat{f}_i$ remains unbiased.

$\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^B \hat{f}_i(X)\right] - \theta = 0.$

T. The variance of am average of B unbiased estimators, $(*)$

which have small covariances ≈ 0, and similar variances $\approx \sigma^2$ my reduce the variance to σ^2/B.

$$\mathrm{Var}\left[\frac{1}{B}\sum_{i=1}^{B}\hat{f}_i(X)\right] = \frac{1}{B^2}\sum_{i=1}^{B}\mathrm{Var}\left[\hat{f}_i(X)\right] + \frac{1}{B^2}\sum_{\substack{i,j\\ i\neq j}}^{B}\mathrm{Cov}\left(\hat{f}_i(X),\hat{f}_i(X)\right) \overset{(*)}{\approx} \frac{\sigma^2}{B}.$$

5.6 — Bias Variance for Squared Loss

$\mathbb{E}_D\left[\mathbb{E}_{\mathbf{X},Y}\left[(Y-\hat{h}_D(\mathbf{X}))^2\right]\right]$ (Note: $h^*(X) = \mathbb{E}_Y[Y\,|\,X]$)

$$= \underbrace{\mathbb{E}_{\mathbf{X}}\left[\left(\mathbb{E}_D\left[\hat{h}_D(\mathbf{X})\right] - h^*(\mathbf{X})\right)^2\right]}_{\text{Bias}^2} + \underbrace{\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[\left(\hat{h}_D(\mathbf{X}) - \mathbb{E}_{D'}\left[\hat{h}_{D'}(\mathbf{X})\right]\right)^2\right]\right]}_{\text{Variance}}$$

$+\mathbb{E}_{\mathbf{X},Y}\left[(Y-h^*(\mathbf{X}))^2\right]$ (Noise)

Derivation: 1) $\pm\mathbb{E}_Y[Y\,|\,X]$ gives noise, 2) Prove vanishing cross-term, 3) $\pm\mathbb{E}_D\left[\hat{f}(X)\right]$ gives bias and variance

5.7 — Gaussian Processes

$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{o},\sigma^2\mathbf{I}_n),\quad \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{o},\boldsymbol{\Lambda}^{-1})$

Since the outputs $\mathbf{y}$ are a linear combination of normally distributed rvs $\boldsymbol{\beta}$, they are jointly Gaussian themselves.

$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\epsilon}}[\mathbf{y}] = \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\epsilon}}[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}] = \mathbf{X}\mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{\beta}] + \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}] = \mathbf{X}\mathbf{o}+\mathbf{o} = \mathbf{o}$.

$\mathrm{Cov}(\mathbf{y}) = \mathbb{E}\left[\mathbf{y}\mathbf{y}^\top\right] = \mathbb{E}\left[(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})^\top\right] = \mathbf{X}\boldsymbol{\Lambda}^{-1}\mathbf{X}^\top + \sigma^2\mathbf{I}$.

Special case: $\Lambda^{-1} := \lambda\mathbf{I}$: $\mathrm{Cov}(\mathbf{y}) = \lambda^{-1}(\mathbf{X}\mathbf{X}^\top + \lambda\sigma^2\mathbf{I}_n)$.

Then we can rewrite the joint distr. as $\mathbf{y} \sim \mathcal{N}(\mathbf{o},\mathbf{K}+\sigma^2\mathbf{I})$

where $k(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i^\top\boldsymbol{\Lambda}^{-1}\mathbf{x}_j$ is a so-called kernel function (could be replaced with others).

Joint D. $P\left(\binom{\mathbf{y}}{y_{n+1}}\,\middle|\,x_{n+1},\mathbf{X},\sigma\right) = \mathcal{N}\left(\binom{\mathbf{y}}{y_{n+1}}\,\middle|\,\mathbf{o},\begin{pmatrix}\mathbf{C}_n & \mathbf{k}\\ \mathbf{k}^\top & c\end{pmatrix}\right)$

$\mathbf{K} = k(\mathbf{X},\mathbf{X})$, $\mathbf{C}_n = \mathbf{K}+\sigma^2\mathbf{I}$, $c = k(x_{n+1},x_{n+1}) + \sigma^2$, $\mathbf{k} = k(x_{n+1},\mathbf{X})$

Pred D. $P(y_{n+1}\,|\,x_{n+1},\mathbf{X},\mathbf{y},\sigma) = \mathcal{N}\left(y_{n+1}\,\middle|\,\boldsymbol{\mu}_{y_{n+1}},\sigma^2_{y_{n+1}}\right)$

$\boldsymbol{\mu}_{y_{n+1}} = \mathbf{k}^\top\mathbf{C}_n^{-1}\mathbf{y} = \mathbf{k}^\top(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y},\ \sigma^2_{y_{n+1}} = c - \mathbf{k}^\top\mathbf{C}_n^{-1}\mathbf{k}$

6 Classification

0/1 Loss $w^* = \arg\min_w \sum_{i=1}^n [y_i \neq sign(w^\top x_i)]$

Perceptron $w^* = \arg\min_w \sum_{i=1}^n [\max(0, y_i w^\top x_i)]$

Exponential Loss $L(y,z) = \exp -(2y-1)(2z-1)$

Logistic Loss $L(y,z) = \ln(1+\exp -(2y-1)(2z-1))$

Dep. on appl. use $\mathbf{y}\mathbf{w}^\top\mathbf{x}$ or $-(2y-1)(2z-1)$ as error

6.1 — Generative VS Discriminative

$$P(Y\,|\,X) = \frac{P(X\,|\,Y)\,P(Y)}{\sum_y P(X\,|\,Y)\,P(Y)}$$

Generative: Est. both $P(Y)$ and $P(X\,|\,Y)$ then use Bayes.

Discr.: Est. $P(Y\,|\,X)$ directly, fitting discr. f. $g(Y,X)$.

6.2 — SVMs

6.2.1 — Primal, constrained

$\min_w w^\top w + C\sum_{i=1}^n \xi_i$, s.t. $y_i w^\top x_i \geq 1-\xi_i, \xi_i \geq 0$

6.2.2 — Primal, unconstrained

$\min_w w^\top w + C\sum_{i=1}^n \max(0, 1-y_i w^\top x_i)$ (hinge loss)

6.2.3 — Dual

$\max_\alpha \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^\top x_j$, s.t. $0 \leq \alpha_i \leq C$

6.2.4 — Dual to Primal

Dual to primal: $w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$, $\alpha_i > 0$: support vector.

6.3 — Kernelized SVMs

$\max_\alpha \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(x_i,x_j)$, s.t. $0 \geq \alpha_i \geq C$

Classify: $y = sign(\sum_{i=1}^n \alpha_i y_i k(x_i,x))$

– How to find a^T?

$a = \{w_0, w\}$ used along $\widetilde{x} = \{1, x\}$

Gradient Descent: $a(k+1) = a(k) - \eta(k)\nabla J(a(k))$

Newton: 2nd ord. Taylor, where $\eta_{opt} = H^{-1}$, $H = \frac{\partial^2 J}{\partial a_i \partial a_j}$

J is the cost mat., popular: Perceptron cost: $J_p(a) = \sum(-a^T\widetilde{x})$

6.4 — Logistic Regression

$P(Y=y\,|\,\mathbf{x},\mathbf{w}) = Ber(y;\sigma(\mathbf{w}^\top\mathbf{x})) = \frac{1}{1+e^{-y\mathbf{w}^\top\mathbf{x}}} = p_y$

$P(Y=+1\,|\,\mathbf{x},\mathbf{w}) = \sigma(\mathbf{w}^\top\mathbf{x}) = p_+$

Gr. st. w. Gau. Pr.: $\mathbf{w} \leftarrow \mathbf{w}(1-2\lambda\eta_t) + \eta_t y\mathbf{x}\hat{P}(Y=-y\,|\,\mathbf{w},\mathbf{x})$

6.4.1 — Multi-Class Logistic Regression

$P(Y=i\,|\,\mathbf{x},\mathbf{w}_1,\ldots,\mathbf{w}_c) = \frac{\exp(\mathbf{w}_i^\top\mathbf{x})}{\sum_{j=1}^c \exp(\mathbf{w}_j^\top\mathbf{x})} = p_i$

7 Multiclass Classification

One-VS-All $y = \arg\max_{i\in\{1,\ldots,c\}} f_i(\mathbf{x})$

One-VS-One $\binom{c}{2}$ bin. clf. for each $(i,j) \in \{1,\ldots,c\}^2$.

$f_{(i,j)}: \mathcal{X} \to \{-1,+1\}$ $y = \arg\max_{i\in\{1,\ldots,c\}} \sum_{j=1 j\neq i}^c \mathbb{1}_{\{f_{(i,j)}(\mathbf{x})=+1\}}$.

8 Jackknife

Method to compensate for syst. est. errors (bias reduction).

Goal: Numerically estimate the bias of an estimator $\hat{S}_n$.

$\tilde{S}_n = \frac{1}{n}\sum_{i=1}^n S_{n-1}^{-i}$ (LOO- Estimator)

Then the est. for the bias of $\hat{S}_n$ is: $\text{bias}^{JK} = (n-1)(\tilde{S}_n - \hat{S}_n)$

Then the unbiased estimate is $\hat{S}^{JK} = \hat{S}_n - \text{bias}^{JK}$.

9 Probabilistic Methods

$$\underbrace{P(\text{model } \theta\,|\,\text{data } D)}_{\text{Posterior}} = \frac{\overbrace{P(\text{data}\,|\,\text{model})}^{\text{Likelihood}} \times \overbrace{P(\text{model})}^{\text{Prior}}}{\underbrace{P(\text{data})}_{\text{Evidence}}}$$

9.0.1 — Maximum (Cond.) LH Est., (MLE)

$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \hat{P}(y_1,\ldots,y_n\,|\,\mathbf{x}_1,\ldots,\mathbf{x}_n,\boldsymbol{\theta})$
$\quad = \arg\min_{\boldsymbol{\theta}} -\sum_{i=1}^n \log\hat{P}(y_i\,|\,\mathbf{x}_i,\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \ldots$ insert, derivate.

9.0.2 — Maximum a Posteriori Estimate, (MAP)

$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}\,|\,D) = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}\,|\,\mathbf{x}_1,\ldots,\mathbf{x}_n,y_1,\ldots,y_n)$
$\quad \overset{\text{i.i.d}}{=} \arg\min_{\boldsymbol{\theta}} -\log P(\boldsymbol{\theta}) - \sum_{i=1}^n \log P(y_i\,|\,\mathbf{x}_i,\boldsymbol{\theta}) = \ldots$ insert, derivate

10 Ensemble Methods

Use combination of simple hyp. (weak lerners) that are sufficiently diverse to prod. a valid sol. with low bias and var.

Bagging: train weak l. on bootstr. sets with equal weights.

Boosting: train on all data, but reweigh misclassified samples higher and use error-sensitive reweighting of classifiers.

Left out: Decision trees (Stump: $h(x) = sign(ax_i - t)$), Random forest: bagging of trees.

10.0.1 — Ada Boost

$f^*(x) = \arg\min_{f\in F}\sum_{i=1}^n \exp(-y_i f(x_i))$ (eff. mins. exp. l.)

Algorithm 1: AdaBoost Algorithm

Initialize the observation weights: $\forall i$: $w_i^{(1)} \leftarrow \frac{1}{n}$
for $b \leftarrow 1$ **to** B **do**
 Fit a clf. $c_b(x)$ to the weigthed training data (acc. to w_i)
 $\epsilon_b \leftarrow \frac{\sum_{i=1}^n w_i^{(b)}\mathbb{1}_{\{c_b(x_i)\neq y_i\}}}{\sum_{i=1}^n w_i^{(b)}}$ (error of the weak learner)
 $\alpha_b \leftarrow \log\left(\frac{1-\epsilon_b}{\epsilon_b}\right)$ (weigh classifier acc. to accuracy)
 Update the datapoint weights for the next step:
 For all $i \in \{1,\ldots n\}$ do
 $w_i^{(b+1)} \leftarrow w_i^{(b)}\exp\left(\alpha_b\mathbb{1}_{\{y_i\neq c_b(x_i)\}}\right)$
return $\hat{c}_B(x) = \mathrm{sign}\left(\sum_{b=1}^B \alpha_b c_b(x)\right)$

Additive log. reg., Bayesian approach (assumes poster.), Newtonlike updates (GD), if prev. clf. bad, next has high weight.

11 Generative Methods

11.1 — Naive Bayes

Features indep. $P(y|x) = \frac{1}{Z}P(y)P(x|y), Z = \sum_y P(y)P(x|y)$

$y = \arg\max_{y'} P(y'|x) = \arg\max_{y'}\hat{P}(y')\prod_{i=1}^d \hat{P}(x_i|y')$

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

11.2 — Fischer's LDA

$J(\mathbf{w}) = \frac{\|\mathbf{w}^\top(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\|}{\mathbf{w}^\top(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)\mathbf{w}}$ $\hat{\mathbf{w}} = (\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ (unscaled)

However, all samples influence boundary (better: points at border, SVM)

12 Unsupervised Learning

12.1 — Gaussian Mixture Modeling

$(\mu^*,\Sigma^*,\pi^*) = \arg\min -\sum_i log\sum_{j=1}^k \pi_j\mathcal{N}(x_i|\mu_i,\Sigma_j)$

12.2 — EM Algorithm

Problem: sum within log-term of likelihood.

E-step: expectation: pick clusters for points.

Calculate $\gamma_j^{(t)}(x_i) = \frac{P(c|\theta^j)(x_i|c,\theta^j)}{\sum P(x_i|\theta)}$ for each i and j

M-Step: maximum LH: adjust clusters to best fit points.

$\text{prior}_j^{(t)} = \pi_j^{(t)} \leftarrow \frac{1}{n}\sum_{i=1}^n \gamma_j^{(t)}(x_i)$

$\mu_j^{(t)} \leftarrow \frac{\sum_{i=1}^n \gamma_j^{(t)}(x_i)(x_i)}{\sum_{i=1}^n \gamma^{(t)}(x_i)}$, $\Sigma_j^{(t)} \leftarrow \frac{\sum_{i=1}^n \gamma_j^{(t)}(x_i)(x_i-\mu_j^{(t)})(x_i-\mu_j^{(t)})^\top}{\sum_{i=1}^n \gamma_j^{(t)}(x_i)}$