Linear Classifiers

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

Decision boundary $H = \{x \in \mathbb{R}^d : \theta \cdot x + \theta_0 = 0\}$

normal to θ so $\theta(x - \tilde{x}) = 0$, θ_0 determines offset $-\theta_0/\|\theta\|$

Can augment x and θ with 1 and θ_0 to have $f(x) = \theta \cdot x$, $H = \{x : \theta \cdot x = 0\}$

Perceptron algorithm

Input: $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{\pm 1\}$

while some $y^i \neq \text{sign}(\theta \cdot x^i)$
pick some misclassified (x^i, y^i)
$\theta \leftarrow \theta + y^i x^i$ (update θ)
return θ

Convergence Theorem
Given linearly separable data for any choice of updates, perceptron algorithm terminates with all data correctly classified. No more than R^2/γ^2 updates

(radius) $R = \max \|x^i\|$ (margin) $\gamma = \min \frac{y^i(\theta \cdot x^i)}{\|\theta\|}$

Since $\theta = \sum_i \alpha^i y^i x^i$, $\theta \cdot x = \sum_i \alpha^i y^i x^i \cdot x$ so we can work with data in any inner product space

View as stochastic gradient descent, where $J(\theta) = \sum_i (-y^i(\theta \cdot x^i))_+$
so $\nabla J(\theta) = \begin{cases} -y^i x^i & \text{if } y^i(\theta \cdot x^i) < 0 \\ 0 & \text{otherwise} \end{cases}$ and update by $\theta \leftarrow \theta - \nabla J(\theta)$

Converges only if data linearly separable, time to converge depends on margin. Solution depends on the starting point

Support Vector Machines

Optimize linear classification by choosing the classifier maximizing $\frac{y^i(\theta \cdot x^i)}{\|\theta\|}$

Find separating θ, $\theta \in y^i \theta \cdot x^i$ and scale θ by $1/\delta$ → $\min \frac{y^i(\theta \cdot x^i)}{\|\theta\|} = \frac{1}{\|\theta\|}$

$\max \frac{1}{\|\theta\|} \to \min_\theta \|\theta\|^2$ s.t. $y^i \theta \cdot x^i \geq 1$ (quadratic program)

Support vectors are the points that satisfy constraints (compressed)
$\theta = \sum_{i=1}^{n} \alpha^i y^i x^i$ ($\alpha^i \neq 0$ for SVs), $\|\theta\|^2 = \sum_{i,j} \alpha^i \alpha^j y^i y^j x^i \cdot x^j$ (just inner products)

Hard Margin SVM
$\min_\theta \|\theta\|^2$ s.t. $y^i \theta \cdot x^i \geq 1$ ($i = 1, \ldots, n$)

Soft Margin SVM
Relax inequalities $y^i \theta \cdot x^i \geq 1$ by introducing slack variables ξ_i
$\min_\theta \|\theta\|^2 + C \sum_{i=1}^{n} \xi_i$ s.t. $y^i \theta \cdot x^i \geq 1 - \xi_i$, $\xi_i \geq 0$ ($\xi_i = (1 - y^i \theta \cdot x^i)_+$)
$\to \min_\theta \|\theta\|^2 + C \sum_{i=1}^{n} (1 - y^i \theta \cdot x^i)_+$

Regularization: small C → small $\|\theta\|^2$, large C → small misclassification

Feature Selection

Linear classifier: $f(x) = \theta \cdot x$, quadratic: $f(x) = \|x - c\|^2 - r^2 = \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_1 \\ x_2 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ -2c_1 \\ -2c_2 \\ c_1^2 + c_2^2 - r^2 \end{pmatrix}$

Quadratic is simply the linear classifier with features $\Phi(x)$
Want to balance complexity of feature with degree of fit

Kernels

We can work in inner product space, so use $k(x_i, x_j) = x_i \cdot x_j$ or $\Phi(x_i) \cdot \Phi(x_j)$
Solve optimization problem with $k(x_i, x_j)$ to determine α_i for $\theta = \sum \alpha^i y^i \Phi(x^i)$
Since we can then compute classification by $\theta \cdot \Phi(x) = \sum \alpha^i y^i (\Phi(x^i) \cdot \Phi(x))$
Degree-m polynomial kernel given by $K_m(x, \tilde{x}) = (1 + x \cdot \tilde{x})^m$
Radial basis function kernel given by $K_{rbf}(x, \tilde{x}) = \exp(-\gamma \|x - \tilde{x}\|^2)$
Scaling parameter γ affects smoothness. small → large many → few points influence
Kernels add modularity to classifier training since same optimization procedure can be used ($k_{ij} = k(x_i, x_j)$ needed)
A degree-m polynomial on $\mathbb{R}^d$ is inner product with $\binom{m+d}{m}$ features

Decision Theory

The Prediction Problem: Given a training set $(x_1, y_1), \ldots, (x_n, y_n)$ choose a function $f: X \to Y$ so that for subsequent (x, y) $f(x)$ is a good predictor of y
Loss function $\ell: Y \times Y \to \mathbb{R}$ where $\ell(\hat{y}, y)$ is cost of predicting $\hat{y}$ as y
Can define identical or asymmetric loss functions
① Assume that (x_i, y_i) and (x, y) are chosen iid from same p.d.f. $X \times Y$
② A good prediction → small expected loss, choose f to minimize $R(f) = \mathbb{E}\,\ell(f(X), Y)$
Risk is misclassification probability: $R(f) = \mathbb{E}\,\ell(f(X), Y) = \mathbb{E}\,\mathbb{1}[f(X) \neq Y] = \Pr(f(X) \neq Y)$
Two class $R(f) = \mathbb{E}\mathbb{E}[\ell(f(X), Y)|X] = \mathbb{E}[\ell(f(X), -1)P(Y=-1|X) + \ell(f(X), 1)P(Y=1|X)] = \mathbb{E}[\mathbb{1}[f(X)=-1]P(Y=1|X) + \mathbb{1}[f(X)=1]P(Y=-1|X)]$
Bayes Decision Rule $f^*(X) = \begin{cases} 1 & \text{if } P(Y=1|X) > P(Y=-1|X) \\ -1 & \text{otherwise} \end{cases}$ $R(\delta) = \sum_{j} \lambda_{ij} P(w_j|x)$
Risk is minimized when $f = f^*$ (where optimal(Bayes) risk $R^* = \inf_f R(f) = R(f^*)$)
Excess Risk $R(f) - R^* = \mathbb{E}[\mathbb{1}[f(X) \neq f^*(X)] |P(Y=1|X) - P(Y=-1|X)|] = \mathbb{E}[\mathbb{1}[f(X) \neq f^*(X)] |2P(Y=1|X) - 1|]$
Excess risk of decision rule can be quantified in terms of a certain distance from f^*

Risk in Regression
Risk is expected squared error $R(f) = \mathbb{E}\,\ell(f(X), Y) = \mathbb{E}(f(X) - Y)^2 = \mathbb{E}\mathbb{E}[(f(X) - Y)^2|X]$
Minimize the conditional expectation of the loss $\mathbb{E}[(f(X) - Y)^2|X]$ by $f^*(X) = \mathbb{E}[Y|X]$

Bias-Variance Decomposition
$R(f) = \mathbb{E}(f(X) - Y)^2 = \mathbb{E}\mathbb{E}[(f(X) - Y)^2|X] = \mathbb{E}\mathbb{E}[(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y)^2|X]$
$= \mathbb{E}\mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2 + 2(\ldots)(\ldots) + (\mathbb{E}[Y|X] - Y)^2|X] = \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2]$
Minimizers $f^*(X) = \mathbb{E}[Y|X]$ so $R(f) = \underbrace{\mathbb{E}[(f(X) - f^*(X))^2]}_{\text{bias}^2} + \underbrace{\mathbb{E}[(f^*(X) - Y)^2]}_{\text{variance}}$ → $R(f) - R^* = \mathbb{E}[(f(X) - f^*(X))^2]$
Using random training data to choose $\hat{f}$, we would like $\mathbb{E}(R(\hat{f}))$ to be small
$\mathbb{E}R(\hat{f}) - R^* = \mathbb{E}[(\hat{f}(X) - f^*(X))^2] = \mathbb{E}[(\hat{f}(X) - \mathbb{E}\hat{f}(X) + \mathbb{E}\hat{f}(X) - f^*(X))^2] = \underbrace{\mathbb{E}[(\hat{f}(X) - \mathbb{E}\hat{f}(X))^2]}_{\text{variance}} + \underbrace{\mathbb{E}[(\mathbb{E}\hat{f}(X) - f^*(X))^2]}_{\text{bias}^2}$

Three approaches to choosing classifiers
① Estimate a generative model by $P(Y)$ and $P(X|Y)$, use Bayes theorem
$P(Y=-1|X) = \frac{P(X|Y=-1)P(Y=-1)}{P(X|Y=-1)P(Y=-1) + P(X|Y=1)P(Y=1)}$ and define plug-in estimator for discriminative model
② Estimate a discriminative model by $P(Y|X)$ and use $\hat{P}(Y|X)$ as actual $P(Y|X)$
Use as plug-in estimator (but only require accuracy around $P(Y=1|X) = 1/2$)
③ Choose a classifier (perceptron, SVM, ...) based on optimization of criterion

Generative and Discriminative Models

Recall $\Pr(A) = \mathbb{E}\Pr(A|X) \cdot \mathbb{E}Y = \mathbb{E}\mathbb{E}(Y|X) \cdot \Pr(A) = \sum \Pr(A \cap B_i) = \sum \Pr(A|B_i)\Pr(B_i) = \mathbb{E}\Pr(A|X)$
$\mathbb{1}_E = \begin{cases} 1 & \text{when } E \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$ so $\Pr(E) = \mathbb{E}\mathbb{1}_E$ and $\mathbb{E}(\mathbb{1}_E) = \mathbb{E}(\mathbb{E}(\mathbb{1}_E|X))$, and $Y = \sum_y y \mathbb{1}[Y=y]$
$\mathbb{E}Y = \mathbb{E}\sum_y y \mathbb{1}[Y=y] = \sum_y y \mathbb{E}\mathbb{1}[Y=y] = \sum_y y \mathbb{E}\mathbb{E}(\mathbb{1}[Y=y]|X) = \mathbb{E}(\sum_y y \mathbb{E}(\mathbb{1}[Y=y]|X)) = \mathbb{E}\mathbb{E}(Y|X)$
For Gaussian class conditional densities $P(X|Y=1)$, $P(X|Y=-1)$ (with the same variance), the posterior probability is logistic: $P(Y=1|X) = \frac{1}{1 + \exp(-x \cdot \theta - \theta_0)}$
Suppose class conditional distributions are Gaussian:
$p(X|Y=-1) = N(\mu_-, \Sigma)$ $P(X|Y=+1) = N(\mu_+, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x - \mu_+)^T \Sigma^{-1}(x - \mu_+))$
(class $+1, -1$ with μ_+/μ_- and Σ)
Consider the 1-dimensional case $X, \mu_+, \mu_- \in \mathbb{R}$, $\Sigma = \sigma^2$, $P(X|Y=+1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x - \mu_+)^2}{2\sigma^2})$
(apply Bayes rule) $P(Y=+1|X) = \frac{P(X|+1)P(+1)}{P(X|+1)P(+1) + P(X|-1)P(-1)} = \frac{P(+1)\exp(-\frac{(x-\mu_+)^2}{2\sigma^2})}{P(+1)\exp(-\frac{(x-\mu_+)^2}{2\sigma^2}) + P(-1)\exp(-\frac{(x-\mu_-)^2}{2\sigma^2})}$
(Simplify and substitute) $P(Y=+1|X) = \frac{1}{1 + \frac{P(-1)}{P(+1)}\exp(\frac{(x-\mu_+)^2 - (x-\mu_-)^2}{2\sigma^2})} \Rightarrow \exp(\frac{\mu_- - \mu_+}{\sigma^2} x + \frac{\mu_+^2 - \mu_-^2}{2\sigma^2} + \log(\frac{P(-1)}{P(+1)}))$
$\exp(-\theta \cdot x - \theta_0)$ $\theta = \frac{\mu_+ - \mu_-}{\sigma^2}$, $\theta_0 = \frac{\mu_-^2 - \mu_+^2}{2\sigma^2} - \log(\frac{P(-1)}{P(+1)})$
In d-dimensions $\theta = \Sigma^{-1}(\mu_+ - \mu_-)$ $\theta_0 = \frac{\mu_-^T \Sigma^{-1} \mu_- - \mu_+^T \Sigma^{-1} \mu_+}{2} - \log \frac{P(-1)}{P(+1)}$
The discriminant direction minimizes Gaussian data overlap
LDA
$p(X|Y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1))$ $\ell = P(X|Y=1)P(Y=1)$
$\log \ell = -\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1) = -\frac{1}{2}x^T \Sigma^{-1} x + \frac{1}{2}x^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_1^T \Sigma^{-1} x - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \log \pi_k$
$\delta_k(x) = \mu_k^T \Sigma^{-1} x - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$ (predict class k that maximizes $\delta_k(x)$)
Assume class conditional distributions as Gaussian, common covariance
MLE estimate $\hat{\pi}_k = \Pr(Y=k) = \frac{n_k}{n}$ $\hat{\mu}_k = \mathbb{E}[X|Y=k] = \frac{1}{n_k}\sum x_i$ $\hat{\Sigma} = \text{Var}[X|Y=k] = \frac{1}{n}\sum_k \sum_{y_i=k}(x_i - \mu_k)(x_i - \mu_k)^T$

Parameter Estimation Methods

We are interested in estimating Bernoulli and Gaussian parameters → classification
Bernoulli Random Variable (biased coin p)
Method of Moments
Choose p such that distribution has same expectation as average of data
Seeing n tosses, define $+1$'s in $x_1 \ldots x_n$ and choose $p = \frac{1}{n}\sum_{i=1}^{n} x_i$
Maximum Likelihood
Count number of $+1$ outcomes and write $\ell(p) = p^k(1-p)^{n-k}$ → log likelihood
$\log \Pr(k \text{ of } n) = \log\binom{n}{k} + k \log p + (n-k)\log(1-p) \to \ell'(p) = \frac{k}{p} - \frac{n-k}{1-p} \to \hat{p} = k/n$
Penalized Maximum Likelihood
Incorporate prior information that p close to some $\hat{p} = \frac{k+1}{n+2}$
$\log(k \text{ of } n) + \log(p(1-p)) = \log\binom{n}{k} + k \log p + (n-k)\log(1-p) + \log(p(1-p)) \to \ell'(p) = \frac{k}{p} - \frac{n-k}{1-p} + \frac{1-2p}{p(1-p)}$
Bayesian estimation
Model p as a random variable and belief about p captured by pr distr over possible values
Can start by modeling as uniform distribution if no a priori preference
Update belief using Bayes Theorem & observation $P(p|X_1 = 1) \propto P(X_1 = 1|p)\pi(p)$ normalize by $\int P(X_1 = 1|q)\,d\pi(p)$
Bayesian approach gives us distribution and uncertainty info
Assumes that parameters randomly chosen with fixed, known distribution
so everything is just RV with known distribution
Bayesian estimation is just a computation of conditional probability distribution
Maximum a priori (MAP) estimate is just mode of the distribution
(max likelihood if uniform prior, penalized MLE with other prior)
Gaussian Random Variable
$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ and would like to estimate $\theta = (\mu, \sigma^2)$ from data
Method of Moments
Choose $\theta = (\mu, \sigma^2)$ to match first moment ($\mathbb{E}X$) and second moment ($\mathbb{E}X^2$) of data
Second moment given by $\mathbb{E}(X-\mu)^2 + \mu^2 = \sigma^2 + \mu^2$ so set $\sigma^2 = \frac{1}{n}\sum x_i^2 - \mu^2$
If we have p parameters, can solve p equations and match corresp number of moments
Maximum Likelihood
Choose $\theta = (\mu, \sigma^2)$ such that distribution gives observed data high prob
$\ell(\mu, \sigma^2) = \log(\prod_{i=1}^{n}(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp(-\frac{(x_i-\mu)^2}{2\sigma^2}))) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$
$\nabla_\mu(\ell) \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\nabla_\sigma(\ell) \to \frac{n}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 \to \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
Yields the same result as moments
Penalized Maximum Likelihood
$\ell(\mu, \sigma^2) - \text{penalty}(\mu, \sigma^2) = \log(\prod_{i=1}^{n}(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp(-\frac{(x_i-\mu)^2}{2\sigma^2}))) - \text{penalty}(\mu, \sigma^2)$
Bayesian Estimates
Prior: $\pi(\theta) = 1$, Posterior: $p(\theta|X_1 = x_1) \propto p(X_1 = x_1|\theta)\pi(\theta)$
Penalized max likelihood and Bayesian estimates effective in high-dimensional settings when number of parameters is high vs. amt data

Multivariate Normal Distribution

Review the Gaussian density function $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$
$\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$ normalization: $\frac{1}{\sqrt{2\pi\sigma^2}}$, arg of exp: $-\frac{(x-\mu)^2}{2\sigma^2} \leq 0$, concave, max at $x = \mu$
location scale
$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$ $(x \in \mathbb{R}^d)$
$\mu \in \mathbb{R}^d$, $\sigma^2 \in \mathbb{R}^{d \times d}$, normalization: $\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}$, arg of exp $-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \leq 0$ concave, max at μ
location $\mu \in \mathbb{R}^d$: $(x-\mu)$, scale $\Sigma \in S_+^d$: $(x-\mu)^T \Sigma^{-1}(x-\mu)$
Consider two independent Gaussian random variables X and Y:
$p(x, y) = p(x)p(y) = \frac{1}{\sqrt{(2\pi)^2\sigma_x^2\sigma_y^2}} \exp(-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp(-\frac{1}{2}([x-\mu_x \; y-\mu_y]\begin{bmatrix}\sigma_x^2 & 0 \\ 0 & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_x \\ y-\mu_y\end{bmatrix}))$
Covariance Matrices
$\Sigma = \begin{bmatrix} \text{var}(X_1) & \cdots & \text{cov}(X_1, X_n) \\ \vdots & & \vdots \\ \text{cov}(X_n, X_1) & \cdots & \text{var}(X_n) \end{bmatrix}$
• Symmetric: $\Sigma_{ij} = \Sigma_{ji}$ • Non-negative diagonal entries: $\Sigma_{ii} \geq 0$
• Positive semidefinite: for all $v \in \mathbb{R}^d$, $v^T \Sigma v \geq 0$
Diagonal Covariance Matrices $\text{cov}(x_i, x_j) = 0$
Consider $X \sim N(\mu, \Sigma)$ where Σ is diagonal, so components of X uncorrelated
Look at super-level sets of pdf: $\mathcal{E}_r = \{x \in \mathbb{R}^d : (x-\mu)^T \Sigma^{-1}(x-\mu) \leq r^2\}$
$(x-\mu)^T \Sigma^{-1}(x-\mu) \leq r^2 \iff \sum_{i=1}^{d} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \leq r^2$
$\mathcal{E}_r$ thus corresponds to axis-aligned ellipsoids in $\mathbb{R}^d$
If all terms but $x_i = \mu_i$: $(x_i - \mu_i)^2 \leq r^2\sigma_i^2 \iff |x_i - \mu_i| \leq r\sigma_i$
The volume of the ellipsoid $\mathcal{E}_r$ is proportional to $\prod_i \sigma_i = \sqrt{|\Sigma|}$
shows that normalization factor, for fixed r, $\Pr(\mathcal{E}_r)$ is same $\forall \mu, \Sigma$
Non-Diagonal Covariance and Diagonalization
Start with a diagonal Σ and consider transformed vector $Y = \mu_x + G(X - \mu_x)$
where G is matrix corresponding to rotation about origin $\begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix}$
Mean μ unchanged by $\mathbb{E}Y = \mu_x + \mathbb{E}[G(X - \mu_x)] = \mu_x + G(\mathbb{E}[X] - \mu_x) = \mu_x$
Covariance is transformed $\Sigma_Y = \mathbb{E}[(Y-\mu_x)(Y-\mu_x)^T] = \mathbb{E}[G(X-\mu_x)(X-\mu_x)^T G^T] = G\mathbb{E}[(X-\mu_x)(X-\mu_x)^T]G^T = G\Sigma_X G^T$
Eigenvalues and Eigenvectors
For a square matrix A, λ is an eigenvalue and x is eigenvector if $Ax = \lambda x$
where eigenvalues are root of characteristic polynomial $\det(A - \lambda I)$
Spectral Theorem
For a symmetric real matrix $A \in \mathbb{R}^{n \times n}$, we can find n orthonormal eigenvectors of A $(v_1 \ldots v_n)$ and eigenvalues $(\lambda_1 \ldots \lambda_n)$ are real
Can then write $AU = U\Lambda$, $A = U\Lambda U^T$ where
$U = [v_1 \; v_2 \; \cdots \; v_n]$, $\Lambda = \text{Diag}(\begin{bmatrix}\lambda_1 \\ \vdots \\ \lambda_n\end{bmatrix})$
Covariance matrix Σ symmetric so $\Sigma = U\Lambda U^T$ and positive semidefinite
so $\forall \lambda_i \geq 0$: $v_i^T \Sigma v_i = v_i^T U\Lambda U^T v_i = \lambda_i \geq 0$
Precision matrix can be written $\Sigma^{-1} = U\Lambda^{-1}U^T$
$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T U\Lambda^{-1}U^T(x-\mu))$
$U^T(x-\mu)$ computes representation of $x-\mu$ in eigenvector basis of Σ
where the covariance is the diagonal matrix Λ^{-1}
Level sets $\mathcal{E}_r = \{x \in \mathbb{R}^d : (x-\mu)^T \Sigma^{-1}(x-\mu) \leq r^2\}$ corresponds to eigenvector aligned ellipsoids in $\mathbb{R}^d$ $(x-\mu)^T \Sigma^{-1}(x-\mu) = (x-\mu)^T U\Lambda^{-1}U^T(x-\mu) = \sum_{i=1}^{d} \frac{(v_i^T(x-\mu))^2}{\lambda_i^2}$
so if $v_2^T(x-\mu) = \cdots = v_d^T(x-\mu) = 0$ then $x \in \mathcal{E}_r$ when $(v_1(x-\mu))^2 \leq r^2\lambda_1^2 \iff |v_1(x-\mu)| \leq r\lambda_1$
Volume of ellipsoid $\mathcal{E}_r$ proportional to $\prod_{i=1}^{d} \lambda_i = \sqrt{|\Lambda|} = \sqrt{|U| \cdot |\Lambda| \cdot |U^T|} = \sqrt{|U\Lambda U^T|} = \sqrt{|\Sigma|}$
Properties of Multivariate Gaussians
Probability: $\int_{\mathbb{R}^d} p(x; \mu, \Sigma)\,dx = 1$ Expectation: $\mathbb{E}[X] = \begin{bmatrix}\mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_d)\end{bmatrix} = \int_{\mathbb{R}^d} x\,p(x; \mu, \Sigma)\,dx = \mu$
Covariance: $\text{Cov}[X] = \mathbb{E}[(X-\mu)(X-\mu)^T] = \int_{\mathbb{R}^d}(x-\mu)(x-\mu)^T p(x; \mu, \Sigma)\,dx = \Sigma$
Sums of independent Gaussians $X \sim N(\mu_X, \Sigma_X)$, $Y \sim N(\mu_Y, \Sigma_Y) \to X+Y \sim N(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$
We know $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, $\text{Cov}(X+Y) = \mathbb{E}[(X+Y)(X+Y)^T] = \mathbb{E}[XX^T + XY^T + YX^T + YY^T] = \mathbb{E}[XX^T] + 0 + 0 + \mathbb{E}[YY^T] = \Sigma_X + \Sigma_Y$
Marginals
Write d-dimensional Gaussian X as $X = \begin{bmatrix}Y \\ Z\end{bmatrix}$, $\mu = \begin{bmatrix}\mu_Y \\ \mu_Z\end{bmatrix}$, $\Sigma = \begin{bmatrix}\Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ}\end{bmatrix}$ where $Y \in \mathbb{R}^m$ and $Z \in \mathbb{R}^{d-m}$, then $Y \sim N(\mu_Y, \Sigma_{YY})$
Mean and covariance of Y evident from definitions, show Gaussian by writing density as integral (integrate out Z), complete squares → $N(\mu_Y, \Sigma_{YY})$
Affine Transformations
Given d-dimensional Gaussian $X \sim N(\mu, \Sigma^2)$, matrix $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m \Rightarrow Y = AX + b \sim N(A\mu + b, A\Sigma A^T)$
e.g. $m = d$, A has full rank ⇒ Y is shifted, scaled version of X
• $\mathbb{E}Y = A\mu + b$ • $\text{Cov}(Y) = \mathbb{E}[(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^T] = \mathbb{E}[(A(X-\mu))(A(X-\mu))^T] = A\mathbb{E}[(X-\mu)(X-\mu)^T]A^T = A\Sigma A^T$
Ex: Given d-dimensional standard normal $X \sim N(0, I)$, $Y = AX + \mu \sim N(\mu, AA^T)$
Given d-dimensional normal $X \sim N(\mu, \Sigma)$ Y semi positive definite Σ, $Y = \Sigma^{-1/2}(X - \mu) \sim N(0, I)$

Regression

Linear Regression

Recall quadratic loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, risk $R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}(f(X) - Y)^2$
For each X, minimize $\mathbb{E}[(f(X) - Y)^2|X]$ by $f^*(X) = \mathbb{E}[Y|X]$ (minimizer)
Bias-variance $R(f) = \underbrace{\mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2]}_{\text{bias}^2} + \underbrace{\mathbb{E}[(\mathbb{E}[Y|X] - Y)^2]}_{\text{variance}} = \mathbb{E}[(f(X) - f^*(X))^2] + R(f^*)$
Given $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, linear (affine) prediction rules $F_{lin} = \{x \mapsto x^T\beta + \beta_0 : \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}\}$
① Consider class of linear prediction rules and minimize empirical risk
② Model the process of generating Y_i's as linear function of X_i's plus Gaussian noise. Compute MLE for linear coefficients.
Both equations should arrive at normal equations: the choice of β corresponds to projection onto linear subspace
Empirical Risk Minimization
Risk $R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}(f(X) - Y)^2$ is the expected squared error while
Empirical risk $\hat{R}(f) = \hat{\mathbb{E}}_n \ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2$ is sample average of squared error
Choose linear prediction rule $f \in F_{lin}$ that minimizes empirical risk by $\hat{f} := \arg\min_{f \in F_{lin}} \hat{\mathbb{E}}_n \ell(f(X), Y) = \arg\min_{f \in F_{lin}} \sum_{i=1}^{n}(f(X_i) - Y_i)^2$
Use convention of offset term β_0 and constant 1 component in X
Solve $\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n}(X_i^T\beta - Y_i)^2 = \arg\min_{\beta \in \mathbb{R}^p} \|X\beta - Y\|^2$, $X = \begin{pmatrix}X_1^T \\ X_2^T \\ \vdots \\ X_n^T\end{pmatrix}$, $Y = \begin{pmatrix}Y_1 \\ \vdots \\ Y_n\end{pmatrix}$ $X \in \mathbb{R}^{n \times p}$ $Y \in \mathbb{R}^n$ $n \gg p$
$RSS(\beta) = \frac{1}{2}\|X\beta - Y\|^2 = \frac{1}{2}(X\beta - Y)^T(X\beta - Y) = \frac{1}{2}\beta^T X^T X\beta - Y^T X\beta + \frac{1}{2}Y^T Y$
$\nabla_\beta RSS(\beta) = X^T X\beta - X^T Y$ $\nabla_\beta^2 RSS(\beta) = X^T X \succeq 0$, so function is convex
$X^T X\beta = X^T Y \Rightarrow \hat{\beta} = (X^T X)^{-1}X^T Y$ (normal equations)
View as finding linear combination of columns of $X = (x_1 \; x_2 \cdots x_n)$ to minimize Euclidean distance to $Y \in \mathbb{R}^n$ by $Y - X\beta = Y - \sum_{j=1}^{p}\beta_j x_j$
Optimal approximation $\hat{Y}$ in space spanned by columns of X has error $Y - \hat{Y}$ orthogonal to column space
$(Y - \hat{Y})^T X = 0 \iff X^T(Y - X\beta) = 0 \iff X^T Y = X^T X\beta$
Risk of empirical risk minimizer $\hat{f}$ close to minimal if
① X come from compact set ② Y_i's have tails not too heavy (e.g. sub-Gaussian)
③ $\|\hat{\theta}\|$ is not too large ④ $n \gg p$

Linear Model with Additive Gaussian Noise

Model conditional distribution of $Y|X=x$ as $P(Y|X=x) = N(x^T\beta, \sigma^2)$ or $Y = X^T\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$
Using MLE, $L(\beta) = \prod_{i=1}^{n} p(Y_i|X_i, \beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2) \to \ell(\beta) = (\text{const}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2$
and we get the least squares solution
Show β unbiased ($\mathbb{E}Y = X\beta$, $\text{Cov}(Y) = \sigma^2 I$) by looking at expectation
$\mathbb{E}\hat{\beta} = \mathbb{E}[(X^T X)^{-1}X^T Y] = (X^T X)^{-1}X^T\mathbb{E}(Y) = (X^T X)^{-1}X^T X\beta = \beta$
$\text{Cov}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = \mathbb{E}[((X^T X)^{-1}X^T Y - \beta)((X^T X)^{-1}X^T Y - \beta)^T] = \ldots = \sigma^2(X^T X)^{-1}$
If data generated by linear model with Gaussian noise, we can compute distr of parameter estimates $\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2)$, calculate approximate confidence sets for param $z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{v_j}}$ which is normal and design tests for non-zero values of param

Regularization in Regression

Bias-variance tradeoff $\mathbb{E}R(\hat{f}) - R^* = \mathbb{E}[(\hat{f}(X) - \mathbb{E}\hat{f}(X))^2] + \mathbb{E}[(\mathbb{E}\hat{f}(X) - f^*(X))^2]$ (bias^2)
Use random training points so $R(\hat{f})$ is RV (variance)
Increase in bias decreases variance and overall complexity
Linear considers number of params in β and size of coefficients
Subset Selection
Idea: Consider 2^p subsets of variables, fit linear model, and decide which is best
Decide which model complexity to use with cross validation
We can use greedy computational shortcuts to get approximately best solution
Forward-stepwise selection - iterate through all fittings using 1 param, include in next iteration (only if better)
Backward-stepwise selection - start with all and remove least contributing
Shrinkage Methods
Encourage linear predictor coefficients to be small using penalty
Ridge Regression
Punish using penalty term $\hat{\beta} = \arg\min_\beta(\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2)$ or
forbid using constraint $\hat{\beta} = \arg\min_\beta \sum_{i=1}^{n}(Y_i - X_i^T\beta)^2$ s.t. $\sum_{j=1}^{p}\beta_j^2 \leq B$
Do not penalize β_0, $RSS = \|Y - X\beta\|^2 + \lambda\|\beta\|^2 \to \hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1}X^T Y$
Even if $X^T X$ singular, penalty term produces an inverse
Solution also depends on scaling of covariates (common to standardize)
Lasso
Use L1 norm $\hat{\beta} = \arg\min_\beta(\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|)$, or with constraint $\hat{\beta} = \arg\min_\beta \sum_{i=1}^{n}(Y_i - X_i^T\beta)^2$ s.t. $\sum_{j=1}^{p}|\beta_j| \leq B$
Ridge regression leads to reduced coefficients, Lasso pushes to zero

Bayesian View of Linear Regression

Given linear model $P(Y|X=x) = N(x^T\beta, \sigma^2)$
Model β as RV ($\beta = N(0, \tau^2 I)$) then compute posterior distribution
$P(\beta|X, Y) \propto P(Y|\beta, X)P(\beta) \to \exp(-\frac{(Y - X^T\beta)^2}{2\sigma^2} - \frac{\|\beta\|^2}{2\tau^2}) = \exp(-\frac{1}{2}(\sum_{i=1}^{n}\frac{(Y_i - X_i^T\beta)^2}{\sigma^2} + \frac{\|\beta\|^2}{\tau^2}))$
MAP estimates β as mode of distribution
Ridge regression ⟺ MAP w/ Gaussian prior, Lasso ⟺ MAP Laplace prior

Logistic Regression

$P(Y=1|x) = \frac{1}{1 + \exp(-\beta^T x)}$ Given data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{0, 1\}$ estimate β by MLE
log likelihood: $\ell(\beta) = \log P(y_1 \ldots y_n|x_1 \ldots x_n; \beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i)\log(1 - \mu_i(\beta))$
where $\mu_i(\beta) = P(Y=1|X=x_i; \beta) = \frac{1}{1 + \exp(-\beta^T x_i)}$, $\nabla_\beta \mu_i(\beta) = \mu_i(\beta)(1 - \mu_i(\beta))x_i$
First derivative: $\nabla_\beta \ell(\beta) = \sum_{i=1}^{n}((\frac{y_i}{\mu_i(\beta)} - \frac{1 - y_i}{1 - \mu_i(\beta)})\nabla_\beta \mu_i(\beta)) = \sum_{i=1}^{n}(y_i - \mu_i(\beta))x_i$
Second derivative: $\nabla_\beta^2 \ell(\beta) = \sum_{i=1}^{n} -\mu_i(\beta)(1 - \mu_i(\beta))x_i x_i^T$
MLE yields $\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \mu_i(\beta)x_i$ • system of p-coupled nonlinear eqs with no closed (general) solution but convex
Gradient Ascent
$\beta^{(t+1)} = \beta^{(t)} + \eta\nabla_\beta \ell(\beta^{(t)}) = \beta^{(t)} + \eta\sum_{i=1}^{n}(y_i - \mu_i(\beta^{(t)}))x_i$
Each gradient calculation takes $O(np)$ time
Stochastic Gradient Descent/Ascent
Instead compute approximation $\approx \nabla_\beta \ell(\beta) = (y_{i_t} - \mu_{i_t}(\beta^{(t)}))x_{i_t}$ which takes $O(p)$ time and update using the equation $\beta^{(t+1)} = \beta^{(t)} + \eta(y_{i_t} - \mu_{i_t}(\beta^{(t)}))x_{i_t}$
Minimizing RSS in linear regression can also be done by gradient/stochastic gradient descent with convex regularizers $\|\beta\|_1, \|\beta\|_2^2$
Newton's Method
Use 1st-order Taylor approximation $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$ and solve for $f(x) = 0$: $x_1 = x_0 - f(x_0)/f'(x_0)$
Solve $\nabla f(x) = 0$ using $x_{t+1} = x_t - [\nabla^2 f(x_t)]^{-1}\nabla f(x_t)$
$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i)\log(1 - \mu_i(\beta))$
$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n}(y_i - \mu_i(\beta))x_i = X^T(Y - \mu)$
$\nabla_\beta^2 \ell(\beta) = \sum_{i=1}^{n} -\mu_i(\beta)(1 - \mu_i(\beta))x_i x_i^T = -X^T \text{diag}(\mu(1-\mu))X$
$\beta^{(t+1)} = \beta^{(t)} + [X^T \text{diag}(\mu(1-\mu))X]^{-1}X^T(Y - \mu)$
Each iteration is least squares with $x_i x_i^T$ weighted by $\mu_i(1 - \mu_i)$ (Iteratively weighted least squares)
View MLE as minimizing log loss for $\hat{p}(y|x) = \begin{cases} f(x) & \text{if } y = 1 \\ 1 - f(x) & \text{if } y = -1 \end{cases}$ Binomial Deviance convex, differentiable, S.D.-hinge loss, similar asymptotes
Define loss as $\ell_{log}(\hat{p}(\cdot|x), y) = -\log(\hat{p}(y|x)) = \begin{cases} -\log f(x) & \text{if } y = 1 \\ -\log(1 - f(x)) & \text{if } y = -1 \end{cases}$
Write logistic log loss $\log(1 + \exp(-y\beta^T x))$
and MLE corresponds to minimizing sample average $\frac{1}{n}\sum \log(1 + \exp(-y_i\beta^T x_i))$

Convex Optimization and SVMs

Recall hard margin SVM: $\min_\theta \|\theta\|^2$ s.t. $y^i\theta \cdot x^i \geq 1$

soft margin SVM: $\min_\theta \|\theta\|^2 + C\sum_{i=1}^{n}(1-y^i\theta\cdot x^i)_+$

AS QP: hard margin SVM: $\min_\theta \|\theta\|^2$ s.t. $y^i\theta\cdot x^i \geq 1$

soft margin SVM: $\min_{\theta,\xi} \|\theta\|^2 + C\sum_{i=1}^{n}\xi_i$ s.t. $\xi_i \geq 0$, $\xi_i \geq 1-y^i\theta\cdot x^i$

Convex sets are those in which it is possible to draw line, all in set

Consider convex optimization problem $p^* = \min_{x\in\mathbb{R}^n} f_0(x)$ s.t. $f_i(x) \leq 0$

Writing constraints as penalties $p^* = \min_{x\in\mathbb{R}^n} f_0(x) + \begin{cases} 0 & \text{if all } f_i(x)\leq 0 \\ +\infty & \text{otherwise}\end{cases}$

Replace constraint penalty with something smaller:

Introduce Lagrange multipliers (dual variables) $\lambda_1,\dots,\lambda_m \geq 0$ and define Lagrangian $L: \mathbb{R}^{n+m}\to\mathbb{R}$ as $L(x,\lambda) = f_0(x) + \sum_{i=1}^{m}\lambda_i f_i(x)$

λ_i cost of violating constraint $f_i(x) \leq 0$

L defines saddle point game: one player (MIN) chooses x to minimize L, the other player (MAX) chooses λ to maximize L. If MIN violates a constraint, $f_i(x) > 0$, MAX can drive L to infinity.

primal problem $p^* = \min_x \max_{\lambda\geq 0} L(x,\lambda)$ (infeasible $x \to \infty$, feasible $\to \lambda_i f_i(x) \to 0$)

dual problem $d^* = \max_{\lambda\geq 0} g(\lambda) = \max_{\lambda\geq 0}\min_x L(x,\lambda)$ $\quad g(\lambda) := \min_x L(x,\lambda)$

This is a zero sum game, always better to play second:

$p^* = \min_x \max_{\lambda\geq 0} L(x,\lambda) \geq \max_{\lambda\geq 0}\min_x L(x,\lambda) = d^*$ (weak duality)

If there is saddle point (x^*,λ^*) so that for all x and $\lambda\geq 0$, $L(x^*,\lambda)\leq L(x^*,\lambda^*)\leq L(x,\lambda^*)$

then $p^* = \min_x \max_{\lambda\geq 0} L(x,\lambda) = \max_{\lambda\geq 0}\min_x L(x,\lambda) = d^*$ (strong duality)

Complementary Slackness

If $p^* = d^*$ and we have primal solution x^* and dual solution λ^* then for i^{th} constraint $(f_i(x)\leq 0)$, $\lambda_i^* f_i(x^*) = 0$.

$f_i(x^*) < 0 \Rightarrow \lambda_i = 0$, $\lambda_i > 0 \Rightarrow f_i(x^*) = 0$

As a result, every term in $\sum_{i=1}^{m}\lambda_i^* f_i(x^*) \geq 0$ is zero

Karush-Kuhn-Tucker Optimality Conditions

Suppose f_0, f_i are convex and differentiable. Then x and λ optimal if and only if:

① Primal feasibility: $f_i(x) \leq 0$ ② Dual feasibility: $\lambda_i \geq 0$

③ Complementary slackness: $\lambda_i f_i(x) = 0$ ④ Stationarity: $\nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) = 0$

SVMs

Hard Margin SVMs $\quad g(\alpha) = \min_\theta L(\theta,\alpha)$

Write $\min_\theta \|\theta\|^2$ s.t. $y^i\theta\cdot x^i \geq 1$ as $L(\theta,\alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{n}\alpha_i(1-y_i\theta^T x_i)$

$\nabla_\theta L(\theta,\alpha) = \theta + (-1)\sum_{i=1}^{n}\alpha_i y_i x_i \to \theta^* = \sum_{i=1}^{n}\alpha_i y_i x_i$

$g(\alpha) = \frac{1}{2}\theta^{*T}\theta + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i y_i \theta^* x_i = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j$

If $\exists$ feasible $\theta \Rightarrow$ strong duality $\to \max_{\alpha,\ \alpha\geq 0} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j$

$\alpha_i > 0 \Rightarrow y_i\theta^{*T}x_i = 1$ and $y_i\theta^T x_i > 1 \Rightarrow \alpha_i = 0$

Express solution as kernel $\hat{y}(x) = \mathrm{sign}(\langle\theta,\Phi(x)\rangle) = \mathrm{sign}\left(\sum_{i=1}^{n}\alpha_i y_i\langle\Phi(x_i),\Phi(x)\rangle\right)$

where α solves dual problem $\min_\alpha \frac{1}{2}\alpha^T \mathrm{diag}(y) K \mathrm{diag}(y)\alpha - \alpha^T 1$ s.t. $\alpha\geq 0$ $\quad K(x_i,x)$

Soft Margin SVMs

Write $\min_\theta \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}(1-y^i\theta\cdot x^i)_+$ as $\min_{\theta,\xi}\frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$ $\quad \xi_i\geq 0$, $\xi_i \geq 1-y^i\theta\cdot x^i$

$L(\theta,\xi,\alpha,\lambda) = \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1-y_i\theta^T x_i - \xi_i) - \sum_{i=1}^{n}\lambda_i\xi_i$

$\nabla_\theta L(\theta,\alpha) = \theta + (-1)\sum_{i=1}^{n}\alpha_i y_i x_i \to \theta^* = \sum_{i=1}^{n}\alpha_i y_i x_i$

$\nabla_\xi L(\theta,\xi,\alpha,\lambda) = \frac{C}{n}\sum 1 - \alpha - \xi \to \alpha_i + \xi_i = \frac{C}{n}$ $\quad$ so $g(\alpha,\lambda) = \sum_i\alpha_i - \frac{1}{2}\sum_{i=1}\alpha_i\alpha_j y_i y_j x_i^T x_j$

Dual problem given by $\max_{\alpha,\lambda}\sum_{i=1}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j$

$\alpha_i \geq 0$, $\lambda_i \geq 0$, $\alpha_i + \lambda_i = C/n$

Consequences of complementary slackness: $\alpha_i^*(1-y_i x_i^T\theta^* - \xi_i^*) = 0$

$\alpha_i^* > 0 \Rightarrow y_i x_i^T\theta^* = 1-\xi_i^* \leq 1$ 'support vectors' $\lambda_i^*\xi_i^* = 0$

lie in the wrong side of half space

If $y_i x_i^T\theta^* < 1$, $\xi_i^* > 0$ so $\lambda_i^* = 0$ and $\alpha_i^* = C/n$ so

support vectors in open halfspace have $\alpha_i^* = C/n$