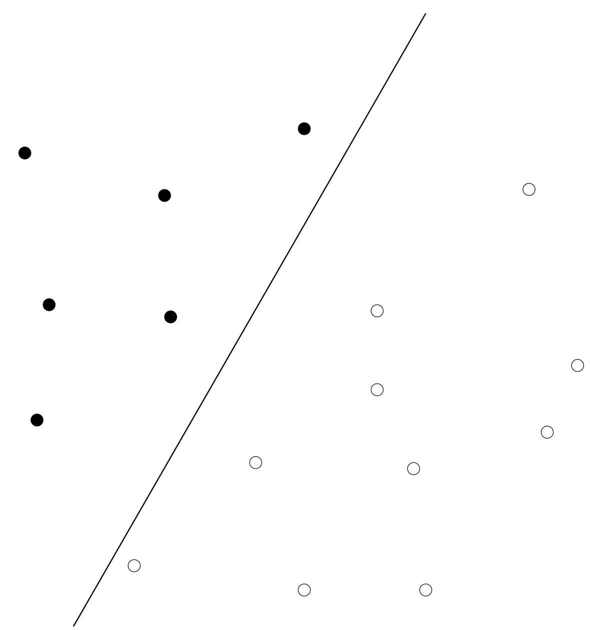
Optimisation Combinatoire et Convexe.

Statistical Applications

Linear discrimination

separate two sets of points $\{x_1, \ldots, x_N\}$, $\{y_1, \ldots, y_M\}$ by a hyperplane:

$$a^T x_i + b > 0, \quad i = 1, \ldots, N, \qquad a^T y_i + b < 0, \quad i = 1, \ldots, M$$

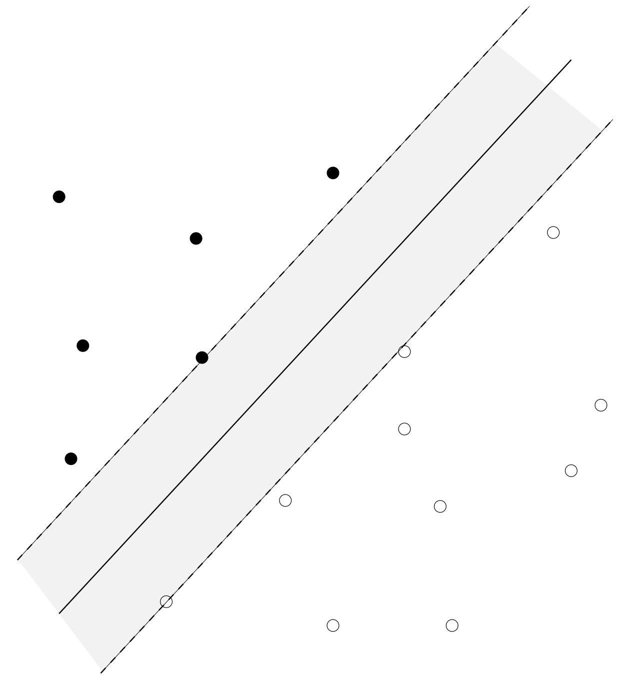


homogeneous in a, b, hence equivalent to

$$a^T x_i + b \geq 1, \quad i = 1, \ldots, N, \qquad a^T y_i + b \leq -1, \quad i = 1, \ldots, M$$

a set of linear inequalities in a, b

Robust linear discrimination



(Euclidean) distance between hyperplanes

$$\begin{aligned} \mathcal{H}_1 &= \{z \mid a^T z + b = 1\} \\ \mathcal{H}_2 &= \{z \mid a^T z + b = -1\} \end{aligned}$$

is $\mathbf{dist}(\mathcal{H}_1, \mathcal{H}_2) = 2/\|a\|_2$

to separate two sets of points by maximum margin,

$$\begin{array}{ll} \text{minimize} & (1/2)\|a\|_2 \\ \text{subject to} & a^T x_i + b \geq 1, \quad i = 1, \ldots, N \\ & a^T y_i + b \leq -1, \quad i = 1, \ldots, M \end{array} \tag{1}$$

(after squaring objective) a QP in a, b

Lagrange dual of maximum margin separation problem

$$
\begin{array}{ll}
\mbox{maximize} & \mathbf{1}^T\lambda + \mathbf{1}^T\mu \\
\mbox{subject to} & 2\left\|\sum_{i=1}^N \lambda_i x_i - \sum_{i=1}^M \mu_i y_i\right\|_2 \leq 1 \\
& \mathbf{1}^T\lambda = \mathbf{1}^T\mu, \quad \lambda \succeq 0, \quad \mu \succeq 0
\end{array} \qquad (2)
$$

from duality, optimal value is inverse of maximum margin of separation

interpretation

- change variables to $\theta_i = \lambda_i/\mathbf{1}^T\lambda$, $\gamma_i = \mu_i/\mathbf{1}^T\mu$, $t = 1/(\mathbf{1}^T\lambda + \mathbf{1}^T\mu)$
- invert objective to minimize $1/(\mathbf{1}^T\lambda + \mathbf{1}^T\mu) = t$

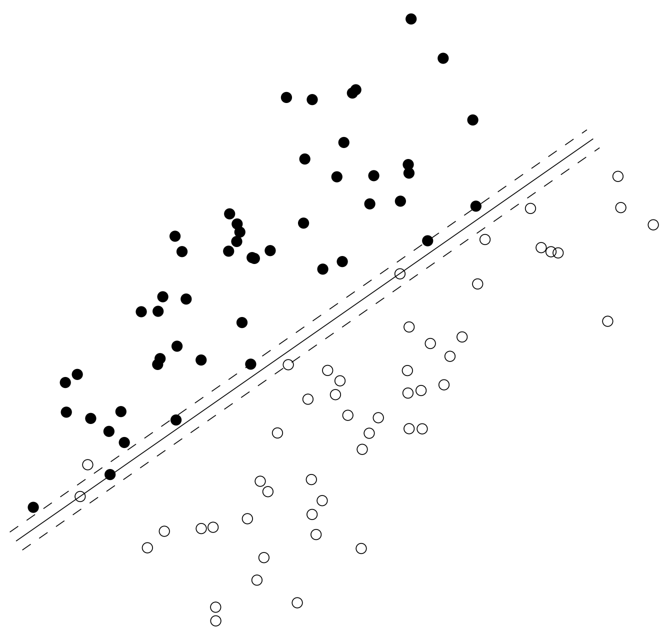
$$
\begin{array}{ll}
\mbox{minimize} & t \\
\mbox{subject to} & \left\|\sum_{i=1}^N \theta_i x_i - \sum_{i=1}^M \gamma_i y_i\right\|_2 \leq t \\
& \theta \succeq 0, \quad \mathbf{1}^T\theta = 1, \quad \gamma \succeq 0, \quad \mathbf{1}^T\gamma = 1
\end{array}
$$

optimal value is distance between convex hulls

Approximate linear separation of non-separable sets

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T u + \mathbf{1}^T v \\ \text{subject to} & a^T x_i + b \geq 1 - u_i, \quad i = 1, \ldots, N \\ & a^T y_i + b \leq -1 + v_i, \quad i = 1, \ldots, M \\ & u \succeq 0, \quad v \succeq 0 \end{array}$$

- an LP in a, b, u, v
- at optimum, $u_i = \max\{0, 1 - a^T x_i - b\}$, $v_i = \max\{0, 1 + a^T y_i + b\}$
- can be interpreted as a heuristic for minimizing #misclassified points

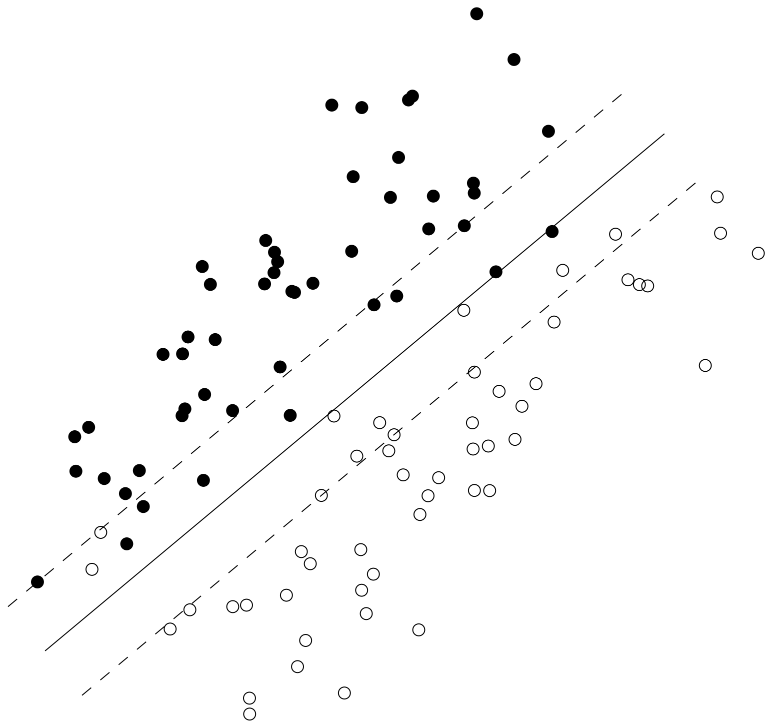


Support vector classifier

$$
\begin{array}{ll}
\text{minimize} & \|a\|_2 + \gamma(\mathbf{1}^T u + \mathbf{1}^T v) \\
\text{subject to} & a^T x_i + b \geq 1 - u_i, \quad i = 1, \ldots, N \\
& a^T y_i + b \leq -1 + v_i, \quad i = 1, \ldots, M \\
& u \succeq 0, \quad v \succeq 0
\end{array}
$$

produces point on trade-off curve between inverse of margin $2/\|a\|_2$ and classification error, measured by total slack $\mathbf{1}^T u + \mathbf{1}^T v$

same example as previous page, with $\gamma = 0.1$:



Support Vector Machines: Duality

Given m data points $x_i \in \mathbb{R}^n$ with labels $y_i \in \{-1, 1\}$.

- The maximum margin classification problem can be written

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|_2^2 + C\mathbf{1}^T z \\ \text{subject to} & y_i(w^T x_i) \geq 1 - z_i, \quad i = 1, \ldots, m \\ & z \geq 0 \end{array}$$

  in the variables w, $z \in \mathbb{R}^n$, with parameter $C > 0$.

- We can set $w = (w, \mathbf{1})$ and increase the problem dimension by 1. So we can assume w.l.o.g. $b = 0$ in the classifier $w^T x_i + b$.
- The Lagrangian is written

$$L(w, z, \alpha) = \frac{1}{2}\|w\|_2^2 + C\mathbf{1}^T z + \sum_{i=1}^{m} \alpha_i (1 - z_i - y_i w^T x_i)$$

  with dual variable $\alpha \in \mathbb{R}_+^m$.

Support Vector Machines: Duality

- The Lagrangian can be rewritten

$$L(w,z,\alpha)=\frac{1}{2}\left(\left\|w-\sum_{i=1}^m \alpha_i y_i x_i\right\|_2^2-\left\|\sum_{i=1}^m \alpha_i y_i x_i\right\|_2^2\right)+(C\mathbf{1}-\alpha)^T z+\mathbf{1}^T\alpha$$

  with dual variable $\alpha\in\mathbb{R}^n_+$.

- Minimizing in (w,z) we form the dual problem

$$\begin{array}{ll}\mbox{maximize} & -\frac{1}{2}\left\|\sum_{i=1}^m \alpha_i y_i x_i\right\|_2^2+\mathbf{1}^T\alpha\\ \mbox{subject to} & 0\leq\alpha\leq C\end{array}$$

- At the optimum, we must have

$$w=\sum_{i=1}^m \alpha_i y_i x_i \quad \mbox{and} \quad \alpha_i=C \mbox{ if } z_i>0$$

  (this is the representer theorem).

Support Vector Machines: the kernel trick

- If we write X the data matrix with columns x_i, the dual can be rewritten

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2}\alpha^T \,\mathbf{diag}(y) X^T X \,\mathbf{diag}(y)\alpha + \mathbf{1}^T\alpha \\ \text{subject to} & 0 \leq \alpha \leq C \end{array}$$

- This means that the data only appears in the dual through the gram matrix

$$K = X^T X$$

  which is called the **kernel** matrix.

- In particular, the original **dimension n does not appear in the dual.** SVM complexity only grows with the number of samples.
- In particular, the x_i are allowed to be infinite dimensional.
- The only requirement on K is that $K \succeq 0$.

Parametric distribution estimation

- distribution estimation problem: estimate probability density $p(y)$ of a random variable from observed values
- parametric distribution estimation: choose from a family of densities $p_x(y)$, indexed by a parameter x

maximum likelihood estimation

$$\text{maximize (over } x\text{)} \quad \log p_x(y)$$

- y is observed value
- $l(x) = \log p_x(y)$ is called log-likelihood function
- can add constraints $x \in C$ explicitly, or define $p_x(y) = 0$ for $x \notin C$
- a convex optimization problem if $\log p_x(y)$ is concave in x for fixed y

Linear measurements with IID noise

linear measurement model

$$y_i = a_i^T x + v_i, \quad i = 1, \ldots, m$$

- $x \in \mathbb{R}^n$ is vector of unknown parameters
- v_i is IID measurement noise, with density $p(z)$
- y_i is measurement: $y \in \mathbb{R}^m$ has density $p_x(y) = \prod_{i=1}^m p(y_i - a_i^T x)$

maximum likelihood estimate: any solution x of

$$\text{maximize} \quad l(x) = \sum_{i=1}^m \log p(y_i - a_i^T x)$$

(y is observed value)

examples

- Gaussian noise $\mathcal{N}(0, \sigma^2)$: $p(z) = (2\pi\sigma^2)^{-1/2} e^{-z^2/(2\sigma^2)}$,

$$l(x) = -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (a_i^T x - y_i)^2$$

  ML estimate is LS solution

- Laplacian noise: $p(z) = (1/(2a)) e^{-|z|/a}$,

$$l(x) = -m \log(2a) - \frac{1}{a} \sum_{i=1}^{m} |a_i^T x - y_i|$$

  ML estimate is ℓ_1-norm solution

- uniform noise on $[-a, a]$:

$$l(x) = \begin{cases} -m \log(2a) & |a_i^T x - y_i| \leq a, \quad i = 1, \ldots, m \\ -\infty & \text{otherwise} \end{cases}$$

  ML estimate is any x with $|a_i^T x - y_i| \leq a$

Logistic regression

random variable $y \in \{0,1\}$ with distribution

$$p = \mathbf{Prob}(y=1) = \frac{\exp(a^T u + b)}{1 + \exp(a^T u + b)}$$

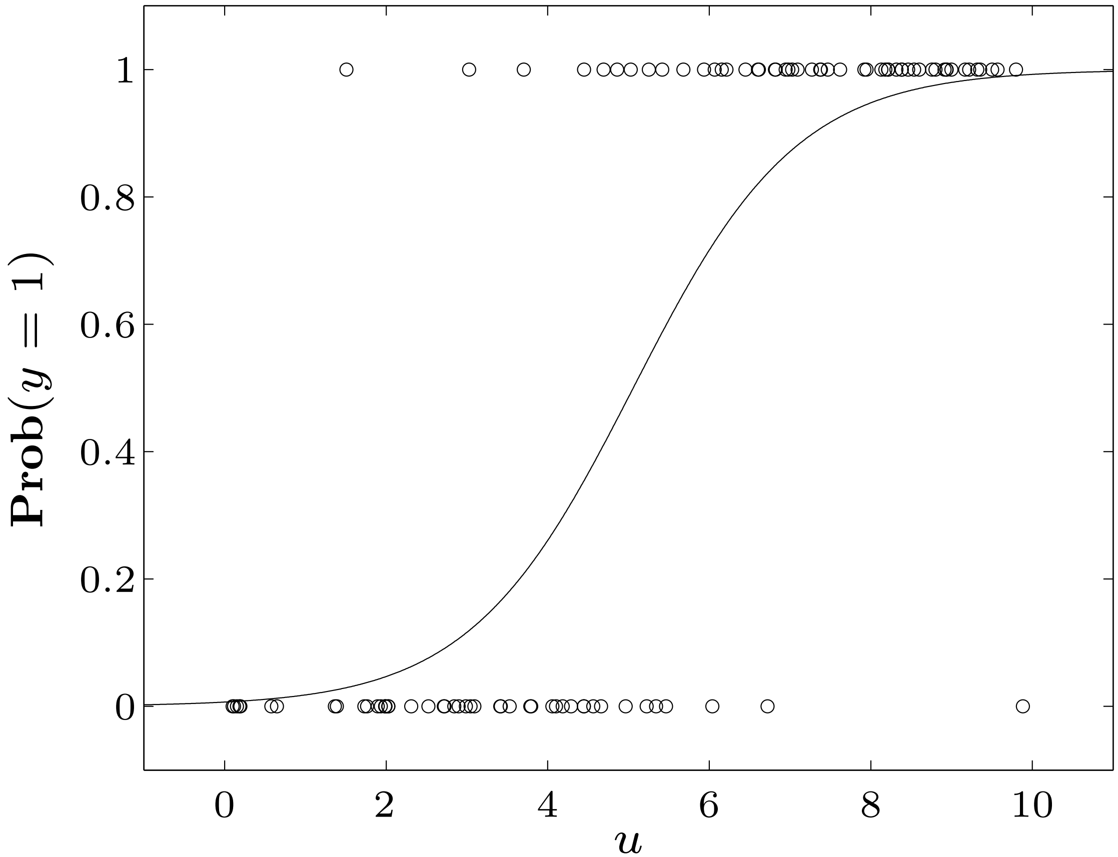
- a, b are parameters; $u \in \mathbb{R}^n$ are (observable) explanatory variables
- estimation problem: estimate a, b from m observations (u_i, y_i)

log-likelihood function (for $y_1 = \cdots = y_k = 1$, $y_{k+1} = \cdots = y_m = 0$):

$$\begin{aligned} l(a,b) &= \log\left(\prod_{i=1}^{k} \frac{\exp(a^T u_i + b)}{1 + \exp(a^T u_i + b)} \prod_{i=k+1}^{m} \frac{1}{1 + \exp(a^T u_i + b)}\right) \\ &= \sum_{i=1}^{k} (a^T u_i + b) - \sum_{i=1}^{m} \log(1 + \exp(a^T u_i + b)) \end{aligned}$$

concave in a, b

example ($n = 1$, $m = 50$ measurements)



- circles show 50 points (u_i, y_i)
- solid curve is ML estimate of $p = \exp(au + b)/(1 + \exp(au + b))$

Experiment design

m linear measurements $y_i = a_i^T x + w_i$, $i = 1, \ldots, m$ of unknown $x \in \mathbb{R}^n$

- measurement errors w_i are IID $\mathcal{N}(0, 1)$
- ML (least-squares) estimate is

$$\hat{x} = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1} \sum_{i=1}^{m} y_i a_i$$

- error $e = \hat{x} - x$ has zero mean and covariance

$$E = \mathbf{E}\, ee^T = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1}$$

  confidence ellipsoids are given by $\{x \mid (x - \hat{x})^T E^{-1} (x - \hat{x}) \leq \beta\}$

experiment design: choose $a_i \in \{v_1, \ldots, v_p\}$ (a set of possible test vectors) to make E 'small'

vector optimization formulation

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = \left(\sum_{k=1}^p m_k v_k v_k^T\right)^{-1} \\ \text{subject to} & m_k \geq 0, \quad m_1 + \cdots + m_p = m \\ & m_k \in \mathbf{Z} \end{array}$$

- variables are m_k (# vectors a_i equal to v_k)
- difficult in general, due to integer constraint

relaxed experiment design

assume $m \gg p$, use $\lambda_k = m_k/m$ as (continuous) real variable

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = (1/m)\left(\sum_{k=1}^p \lambda_k v_k v_k^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1 \end{array}$$

- common scalarizations: minimize $\log\det E$, $\mathbf{Tr}\, E$, $\lambda_{\max}(E)$, . . .
- can add other convex constraints, $e.g.$, bound experiment cost $c^T\lambda \leq B$

Experiment design

D-optimal design

$$\begin{array}{ll} \text{minimize} & \log\det\left(\sum_{k=1}^{p}\lambda_k v_k v_k^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1 \end{array}$$

interpretation: minimizes volume of confidence ellipsoids

dual problem

$$\begin{array}{ll} \text{maximize} & \log\det W + n\log n \\ \text{subject to} & v_k^T W v_k \le 1, \quad k = 1,\ldots,p \end{array}$$

interpretation: $\{x \mid x^T W x \le 1\}$ is minimum volume ellipsoid centered at origin, that includes all test vectors v_k

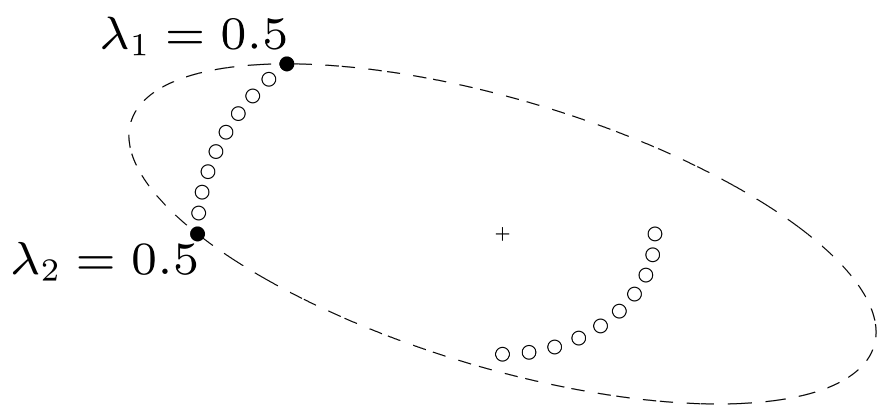
complementary slackness: for λ, W primal and dual optimal

$$\lambda_k(1 - v_k^T W v_k) = 0, \quad k = 1,\ldots,p$$

optimal experiment uses vectors v_k on boundary of ellipsoid defined by W

Experiment design

example $(p = 20)$



design uses two vectors, on boundary of ellipse defined by optimal W

Derivation of dual.

first reformulate primal problem with new variable X

$$\begin{array}{ll}\text{minimize} & \log\det X^{-1} \\ \text{subject to} & X = \sum_{k=1}^{p} \lambda_k v_k v_k^T, \quad \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1\end{array}$$

$$L(X,\lambda,Z,z,\nu) = \log\det X^{-1} + \mathbf{Tr}\left(Z\left(X - \sum_{k=1}^{p}\lambda_k v_k v_k^T\right)\right) - z^T\lambda + \nu(\mathbf{1}^T\lambda - 1)$$

- minimize over X by setting gradient to zero: $-X^{-1} + Z = 0$
- minimum over λ_k is $-\infty$ unless $-v_k^T Z v_k - z_k + \nu = 0$

Dual problem

$$\begin{array}{ll}\text{maximize} & n + \log\det Z - \nu \\ \text{subject to} & v_k^T Z v_k \le \nu, \quad k = 1,\ldots,p\end{array}$$

change variable $W = Z/\nu$, and optimize over ν to get dual of page 17.