Classification by linear methods:

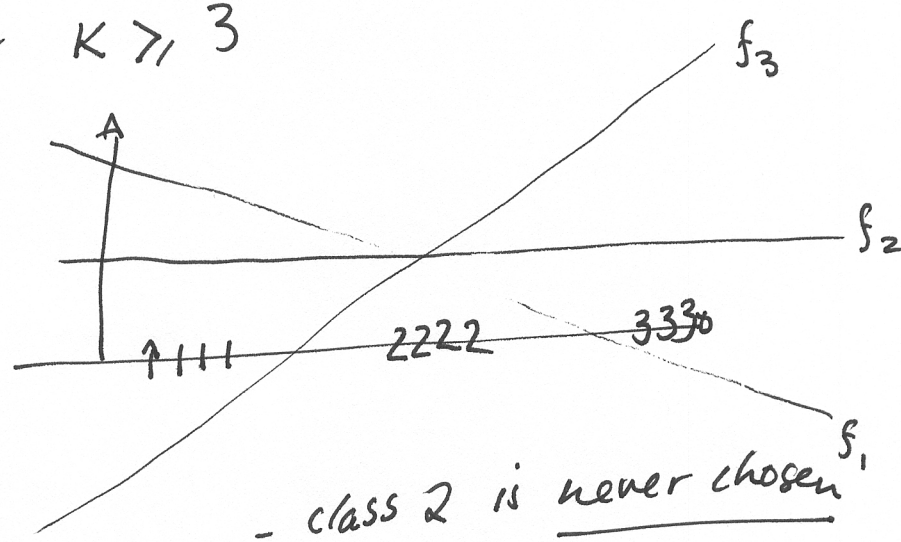
(Ch4 Hastie Tibshirani Friedman).

- k - classes
- generally, fit k fns, and choose class w/ largest
  - discriminant fns.
- Simplest:

train | - use linear regression
| - fit k b.r.s to indicator vars.

test | - eval all k,
| - choose largest.

- Serious problem for $K \geq 3$
  - masking



- class 2 is never chosen

- one might fix this w/ polynomial terms, but this is expensive

Linear Discriminant Analysis

1) Notice that decision theory gives us:

choose k for which $P(k|x)$ is largest

this is $$\frac{P(x|k) \cdot P(k)}{\sum_{i \in \text{classes}} P(x|i) P(i)}.$$

so that a model of $P(x|i)$ is very valuable

2) Assume: $P(x|i) = N(\mu_i, \Sigma)$ ↑ all classes have the same covar

so that $$P(x|i) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu_i)\Sigma^{-1}(x-\mu_i)}$$

3) we can classify with $\log P(k|x)$ because ~~it's~~ log monotone

$$\log\left[\frac{P(k|x)}{P(l|x)}\right] = \log\left[\frac{P(x|k)}{P(x|l)}\right] + \log\left[\frac{\pi_k}{\pi_l}\right]$$

and in our case, this is

$$-\frac{1}{2}\left[(\mu_K+\mu_\ell)'\Sigma^{-1}(\mu_K-\mu_\ell)\right] + x^T\Sigma^{-1}(\mu_K-\mu_\ell) + \log\left(\frac{\pi_K}{\pi_\ell}\right)$$

↑ linear in x (don't forget the priors!)

we can describe this decision rule w/ linear discriminant fns

$$\delta_K(x) = x'\Sigma^{-1}\mu_K - \frac{1}{2}\mu_K'\Sigma^{-1}\mu_K + \log\pi_K$$

classification rule: choose class w/ largest

Note: Σ, & μ_i unknown ∴ must estima

- notice we should have a quite good est for Σ.

Similar trick works for situation where covariance varies ("heteroskedastic")

- but now we have quadratic disc

Quadratic discs

$$\delta_k = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) + \log \pi_k$$

- quadratic decision boundary
- more estimation ⇒ ~~less~~ more variance

Uncomfortable truths:

- linear decision boundaries are v. effective
- LDA is a good way to estimate them
- LDA / QDA are very competitive classifiers

Feature trick:

- a boundary that isn't linear in $\underline{x}$ might be linear in $\underline{\varphi}(\underline{x})$.

Regularization:

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

or even

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\sigma^2 I$$

α, γ, est by cross-validation

Fisher linear discriminants

(or discriminant coordinates or canonical variates)

Find $z = \alpha^T x$ st $\frac{\text{between class var of } z}{\text{within class var of } z}$ is maximised

cf. Principal coordinates

find $z = \alpha^T x$ st variance of z is maximised

between class var $\Sigma_B = \frac{1}{\#\text{ of classes} - 1} \left[\sum_k (\mu_k - \mu_{data})(\mu_k - \mu_{data})^T \right]$

within class var $\Sigma_w = \frac{1}{\#\text{ data} - 1} \left[\sum_{i \in data} [x_i - \mu_{c(i)}][x_i - \mu_{c(i)}]^T \right]$

for $z = \alpha^T x$

$$\text{b.c.v.} = z' \Sigma_B z$$

$$\text{w.c.v.} = z' \Sigma_w z$$

$$\therefore \max \frac{z' \Sigma_B z}{z' \Sigma_w z} \quad \leftarrow \text{hard}$$

$$\max z' \Sigma_B z \quad \text{s.t.} \quad z' \Sigma_w z = 1$$

Logistic regression:

$$\log\left[\frac{P(i|x)}{P(K|x)}\right] = \beta_i' x$$ ← there's a 1 in this.

$$\therefore P(i|x) = \frac{\exp[\beta_i' x]}{1 + \sum_{\ell=1}^{K-1}\left[\exp(\beta_\ell' x)\right]}$$ ← K class

Note:
- this gives a P.L. decision boundary
- if K = # of classes = 2 linear "

Fitting:
- by max likelihood

2 class case:

$$P(1|x) = p(x;\beta) = 1 - p(0|x)$$

$$p(x;\beta) = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)}$$

$$\ell(\beta) = \sum_{i \in \text{data}}\left[\log\left[P(y_i|x_i;\beta\right]\right]$$

$$= \sum_{i \in \text{data}}\left[y_i \log p(x_i;\beta) + (1-y_i)\log[1-p(\right.$$

$$= \sum_i \left[\begin{array}{l} y_i \beta' x \\ \quad - y_i \log(1 + \exp(\beta' x)) \\ \quad - (1-y_i)\log(1 + \exp(\beta' x)) \end{array}\right]$$

Max yields

$$\nabla \ell = 0 = \nabla \left[\sum_i \left\{ y_i \beta' x_i - \log\left(1 + e^{\beta' x_i}\right) \right\} \right]$$

$$= \sum_i x_i \left(y_i - p(x_i; \beta) \right)$$

↑ some mild reorg

and $H_\ell = - \sum_i x_i x_i' \left[p(x_i; \beta)\left(1 - p(x_i; \beta)\right) \right]$

write: $\underline{y} = \text{vec } y_i$; $\underline{\underline{X}} = \begin{bmatrix} x_1' \\ \vdots \\ x_N' \end{bmatrix}$;

$\underline{p}^{(n)} = \text{vec } p(x_i; \beta^{(n)})$

$$\underline{\underline{W}}^{(n)} = \text{diag}\left[p(x_i; \beta^{(n)}) \cdot \left(1 - p(x_i; \beta^{(n)})\right) \right]$$

Newton's method becomes

$$\beta^{(n+1)} = \beta^{(n)} + \underbrace{\left(\underline{\underline{X}}^T \underline{\underline{W}}^{(n)} \underline{\underline{X}}\right)}_{\text{Hessian}}{}^{-1} \underbrace{\underline{\underline{X}}' \left(\underline{y} - \underline{p}^{(n)}\right)}_{\text{gradient}}$$

$$= \left(X^T W^{(n)} X\right)^{-1} \cdot \left[X^T W z\right]$$

where

$$z = X \beta^{(n)} + W^{-1}(y - p)$$

But this is a form of iteratively reweight

least squares.

Problem is convex but th

is mi

- We could focus on search for decision boundary (rather than on prob model)
- e.g. choose best linear boundary

Perceptron

$y_i \in \{1; -1\}$

minimize

$$Err = -\sum_{i \in \text{Misclassified}} y_i (x_i' \beta + \beta_0)$$

$$\nabla E = \begin{bmatrix} -\sum_{i \in M} y_i x_i \\ -\sum_{i \in M} y_i \end{bmatrix}$$

minimize by stochastic gradient descent.

- known to converge in finite # of steps for separable data

* but
  - more than one soln
  - can be many steps
  - if not separable, mischief including long cycles.

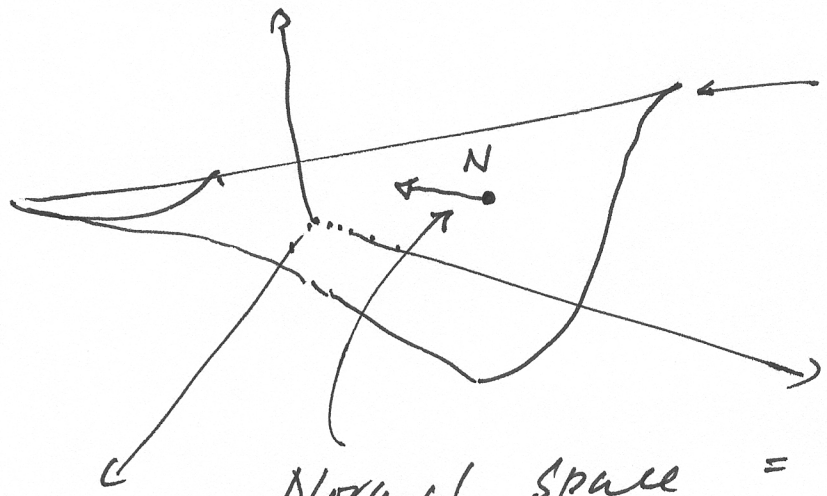
Sideline:

· Constrained optimization

Consider

$$\max \; f_0(x) \quad \text{s.t.} \quad f_1(x) \cdots f_K(x) = 0$$

at an extremal value, ~~∇f~~ ∇f will be normal to constraint surface



Normal space = span $[\nabla f_1 \cdots \nabla f_K]$

$\therefore \nabla f_0 \in \text{span}\left[\nabla f_1 \cdots \nabla f_K\right]$

$\therefore \nabla f_0 + \lambda_1 \nabla f_1 + \lambda_2 \nabla f_2 \cdots \lambda_K \nabla f_K = 0$

and $f_1 = 0$

$f_2 = 0$

$f_3 = 0$

⋮

This is what Lagrangian is all about.

Inequalities:

$\max f_0(x)$ st $f_1(x) \cdots f_K = 0$

$g_1(x) \cdots g_{K'} \geq 0$

if a constraint i is active $\equiv g_i = 0$

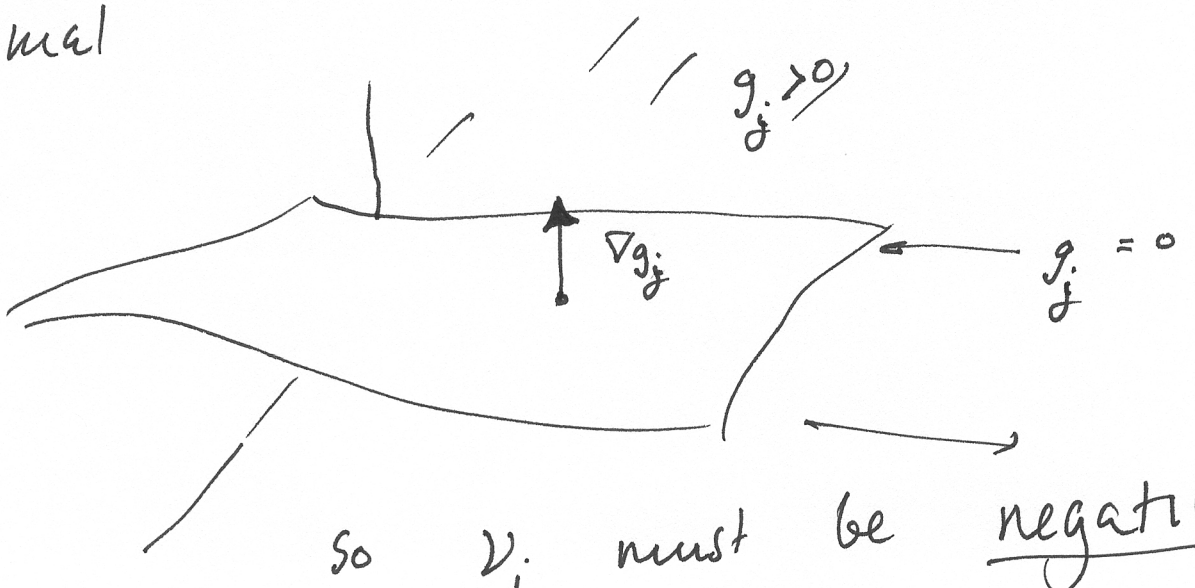
inactive $\equiv g_i > 0$

i active $\Rightarrow \nabla f_0$ is normal to g_i, too

i inactive $\Rightarrow g_i$ irrelevant.

$$\therefore \nabla f_0 + \sum_i \lambda_i \nabla f_i + \sum_{j \in \text{active}} \nu_j \nabla g_j = 0$$

notice that we care about dir of normal



so ν_j must be negative

Whence

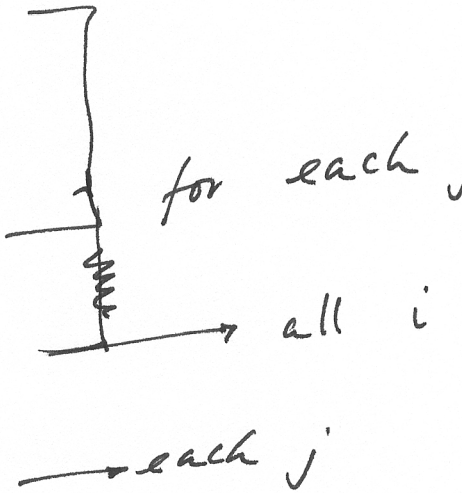
$$\nabla f_0 + \sum_i \lambda_i f_i + \sum_{j \in \text{all}} \nu_j g_j = 0$$

$$\nu_j \leq 0$$

$$\nu_j g_j = 0$$

$$f_i = 0$$

$$g_j \geq 0$$



Karush - Kuhn - Tucker conditions