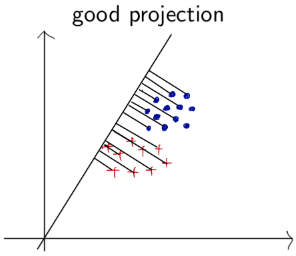
Linear classifiers
Linear discriminant analysis (LDA)
Logistic regression
Support vector machine (SVM)

LDA
Assume classes are multivariate Gaussian with common covariance matrix $\mathbf{\Sigma}$
It can be shown that under these assumptions the decision boundary between any two classes is linear, i.e. a hyperplane in $\mathbb{R}$.

w is a weight vector that defines a projection into a 1-dimensional subspace. Then, in that sub-space, we set a threshold t as the boundary between the two classes.



Fisher's optimal projection: maximize between-class variance relative to within-class variance. In other words: projected class centroids are far apart, projected data are close to centroids.

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmax}} \frac{w'\mathbf{B}w}{w'\mathbf{W}w}$$

$$\mathbf{B} = \sum_c (\boldsymbol{\mu}_c - \bar{\mathbf{x}})(\boldsymbol{\mu}_c - \bar{\mathbf{x}})'$$

$$\mathbf{W} = \sum_c \sum_{i \in c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)'$$

Because w is invariant to rescaling, we can choose w such that $w'\mathbf{W}w = 1$, leading to:

$$w^* = \underset{w}{\operatorname{argmax}}\, w'\mathbf{B}w$$
$$\text{s.t.}$$
$$w'\mathbf{W}w = 1$$

This is a generalized eigenvalue problem, with w given by the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$

If we assume Gaussian class densities and common covariance matrix $\mathbf{\Sigma}$, then LDA is optimal (equivalent to Bayes).

Interlude: Lagrangian optimization
Suppose that we have a standard optimization problem with m inequality and/or p equality constraints:
Minimize $f_0(x)$
s.t. $f_i(x) \leq 0, \quad i = 1...m$
$h_i(x) = 0, \quad i = 1...p$
The Lagrangian is:

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)$$

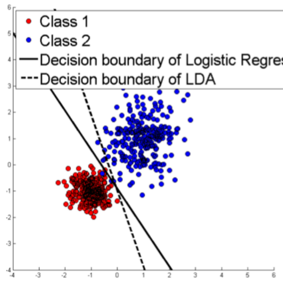
The dual function is the minimum value of the Lagrangian over x:

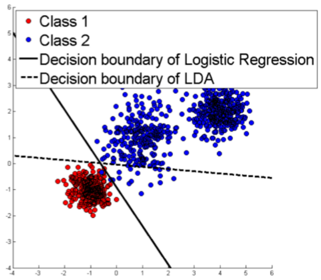
$$g(\lambda, v) = \inf_x L(x, \lambda, v)$$

This function is concave, even if the original problem is not convex. It also is always less than or equal to the original objective function evaluated at its optimal value. Putting these together, if we maximize λ and v in the dual, it is the same as minimizing x in the primal.

Logistic

Same as usual. Estimate using MLE. Can add a regularizer. Often outperforms LDA because it doesn't have squared error (bad). Example:





SVM

Key advantages: no distribution assumptions and data compression (classification only depends on a relatively small subset of overall data).

Hard margin

Data must be linearly separable.

We want to construct a decision hyperplane $H_{w_0+\langle w,x\rangle}$.

Notation: $w_0 + \langle w, x\rangle = \mathbf{w}^T\mathbf{X} + b = \sum w_i x_i + b$. For example, fo line $3x + 2$, $w = 3$ and $w_0 = 2$.

We will require w to be *canonical* to x, i.e.

$$\min_{1\leq i\leq n} |w_0 + \langle w, x_i\rangle| = 1$$

The L2 distance of x to the hyperplane is:

$$\min_{z\in H} ||z - x||_2$$

Which can be shown to be equivalent to:

$$\frac{w_0 + \langle w, x\rangle}{||w||_2}$$

So, the closest point has distance to H of:

$$\frac{1}{||w||_2}$$

We want to maximize the size of the margin, which means maximizing the minimum distance from x to H, which means maximizing the above expression...

$$\max_{w_0,w} \frac{1}{||w||_2}$$

...while keeping it canonical

$$s.t. \quad w_0 + \langle w, x \rangle = 1$$

...and correctly separating the classes

$$\text{sign}(w_0 + \langle w, x \rangle) = Y_i$$

Making the above big is equivalent to making the L2 norm of w small:

$$\min w_0, w\frac{1}{2}||w||_2$$

...and we can cleverly combine the two previous constraints

$$s.t. \quad Y_i(w_0 + \langle w, x \rangle) \geq 1$$

This is the *hard margin* SVM problem in its usual form. It is a nice convex optimization problem that we can solve in the usual way.

Visualization of the hard margin SVM: