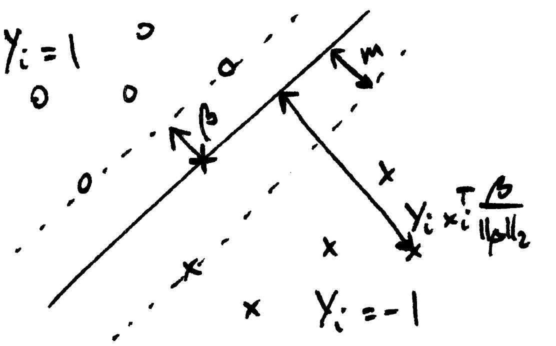
1.6 Other kernel machines

The least squares loss of KRR seems appropriate when the response is cts. We now consider the case where $Y \in \{-1, 1\}^n$, so the responses are class labels.

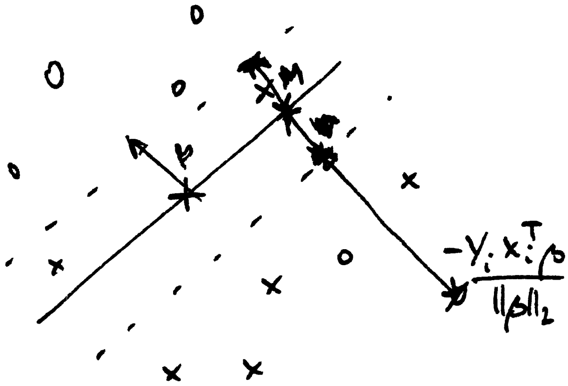
1.6.1 The support vector machine

Suppose $\{x_i\}_{i: Y_i = 1}$ and $\{x_i\}_{i: Y_i = -1}$ are separable by a hyperplane through the origin, i.e. $\exists \beta \in \mathbb{R}^p$ s.t. $Y_i x_i^T \beta > 0 \ \forall i$.



There is an infinite number of planes that separate the two classes. One approach is to pick the plane such that maximises the margin between the two classes. This is given by the optimisation problem

$$\max_{\beta \in \mathbb{R}^p, M \geq 0} M \quad \text{subj to} \quad \frac{Y_i x_i^T \beta}{\|\beta\|_2} \geq M.$$



We can replace the constraint $\frac{Y_i x_i^T \beta}{\|\beta\|_2} \geq M$ with a penalty for how far over its margin boundary x_i is. The penalty should be zero for those points on the correct side of their margin boundary.

Two natural choices for this penalty are

$$\lambda \sum_{i=1}^n \left(M - \frac{Y_i x_i^T \beta}{\|\beta\|_2}\right)_+$$

$$\lambda \sum_{i=1}^n \left(1 - \frac{Y_i x_i^T \beta}{M \|\beta\|_2}\right)_+$$

The second of these leads to a tractable optimisation problem. Replacing $\max M$ with $\min \frac{1}{M^2}$ and adding the penalty, we get

$$\min_{M \geq 0, \beta \in \mathbb{R}^p} \frac{1}{M^2} + \lambda \sum_{i=1}^n \left(1 - \frac{Y_i x_i^T \beta}{M \|\beta\|_2}\right)_+$$

Since the objective function is invariant to β being multiplied by any positive scalar, we can enforce that $\frac{1}{M} = \|\beta\|_2$, thus eliminating M from the objective function. Replacing λ with $\frac{1}{\lambda}$, we arrive at

$$\min_{\beta \in \mathbb{R}^p} \lambda \|\beta\|_2^2 + \sum_{i=1}^n (1 - Y_i x_i^T \beta)_+$$

We have restricted ourselves to using hyperplanes through the origin but more generally we'd like to use translations of these as well. This results in the <u>support vector classifier</u>

$$\underset{(\mu,\beta) \in \mathbb{R} \times \mathbb{R}^p}{\arg\min} \sum_{i=1}^n (1 - Y_i (x_i^T \beta + \mu))_+ + \lambda \|\beta\|_2^2$$

Note that letting $\mathcal{H}$ be the RKHS corresponding to the linear kernel, we can rewrite this as

$$\underset{(\mu,f) \in \mathbb{R} \times \mathcal{H}}{\arg\min} \sum_{i=1}^n (1 - Y_i (f(x_i) + \mu))_+ + \lambda \|f\|_{\mathcal{H}}^2 \qquad (*)$$

The representer theorem (see the variant in Ex 1, Qn 10) tells us that $(*)$ is equiv to

$$\underset{(\mu,\alpha) \in \mathbb{R} \times \mathbb{R}^n}{\arg\min} \sum_{i=1}^n (1 - Y_i (K_i^T \alpha + \mu))_+ + \lambda \alpha^T K \alpha \quad ,$$

which is the <u>support vector machine</u> (SVM). Moreover this equivalence holds for arbitrary RKHS's $\mathcal{H}$ and corresponding kernel matrices K ($K_{ij} = k(x_i, x_j)$).

Predictions at a new x are given by

$$\mathrm{sgn}\left(\hat{\mu} + \sum_{i=1}^n \hat{\alpha}_i k(x, x_i)\right)$$

where $(\hat{\mu}, \hat{\alpha})$ minimises the above.

1.6.2 Logistic regression

Recall that standard logistic regression is motivated by assuming $\log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = -1)}\right) = x_i^T \beta^0$ and that $Y_1, \dots, Y_n$ are independent. An estimate for β^0 is obtained through maximising the likelihood, or equivalently

$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^n \log(1 + \exp(-Y_i x_i^T \beta))$$

The 'kernelised' version is given by $\underset{f \in \mathcal{H}}{\arg\min} \left\{ \sum_{i=1}^n \log(1 + \exp(-Y_i f(x_i))) + \lambda \|f\|_{\mathcal{H}}^2 \right\}$

As in the case for the SVM the representer theorem gives us a finite-dim optim problem equiv to above.