Linear SVM Classifier with slack variables (hinge loss function)

Optimal margin classifier with slack variables and kernel functions described by Support Vector Machine (SVM).

\[
\min_{(w, \xi)} \frac{1}{2}||w||^2 + \gamma \sum \xi(i) \\
\text{subject to } \xi(i) \geq 0 \ \forall i , \ d(i) (w^T x(i) + b) \geq 1 - \xi(i) , \forall i, \text{ and } \gamma > 0.
\]

In dual space

\[
\max W(\alpha) = \sum \alpha(i) - \frac{1}{2} \sum \alpha(i) \alpha(j) \ d(i) \ d(j) \ x(i)^T x(j) \\
\text{subject to } \gamma \geq \alpha(i) \geq 0, \text{ and } \sum \alpha(i) \ d(i) = 0 .
\]

Weights can be found by \( w = \sum \alpha(i) \ d(i) \ x(i) . \)
Solving QP Problem

- Quadratic programming problem with linear inequality constraints.
- Optimization problem involves searching space of feasible solutions (points where inequality constraints satisfied).
- Can solve problem in primal or dual space.
QP software for SVM

- Matlab (easy to use, choose primal or dual space, slow): quadprog()
  - Primal space (w,b, $\xi^+$, $\xi^-$)
  - Dual space ($\alpha$)

- Sequential Minimization Optimization (SMO) (specialized for solving SVM, fast): decomposition method, chunking method

- SVM light (fast): decomposition method
Example

Drawn from Gaussian data $\text{cov}(X) = I$

$20 + \text{pts. Mean} = (.5,.5)$
$20 - \text{pts. Mean} = -(0.5,0.5)$
Primal Space (matlab)
x = randn(40,2);
d = [ones(20,1); -ones(20,1)];
x = x + d * [.5 .5];
H = diag([0 1 1 zeros(1,80)]);
gamma = 1;
f = [zeros(43,1); gamma*ones(40,1)];
Aeq = [d x.*(d*[1 1]) -eye(40) eye(40)];
beq = ones(40,1);
A = zeros(1,83);
b = 0;
lb = [-inf*ones(3,1); zeros(80,1)];
ub = [inf*ones(83,1)];
[w,fval] = quadprog(gamma*H,f,A,b,Aeq,beq,lb,ub);
Example continued

Dual Space (matlab)

\[
\begin{align*}
\text{xn} &= \text{x} \ast (\text{d} \ast [1 \ 1]); \\
\text{k} &= \text{xn} \ast \text{xn}' ; \\
\text{gamma} &= 1; \\
\text{f} &= -\text{ones}(40,1); \\
\text{Aeq} &= \text{d}'; \\
\text{beq} &= 0 \\
\text{A} &= \text{zeros}(1,40); \\
\text{b} &= 0; \\
\text{lb} &= \text{zeros}(40,1); \\
\text{ub} &= \text{gamma} \ast \text{ones}(40,1); \\
[\text{alpha}, \text{fvala}] &= \text{quadprog} (\text{k}, \text{f}, \text{A}, \text{b}, \text{Aeq}, \text{beq}, \text{lb}, \text{ub});
\end{align*}
\]
Example continued

- $w = (1.4245,.4390)^T$ $b = 0.1347$
- $w = \sum \alpha(i) d(i) x(i)$ (26 support vectors, 3 lie on margin hyperplane)
  - $\alpha(i)=0$, $x(i)$ above margin
  - $0 \leq \alpha(i) \leq \gamma$, $x(i)$ lie on margin hyperplanes
  - $\alpha(i) = \gamma$, $x(i)$ lie below margin hyperplanes
- Hyperplane can be represented in
  - Primal space: $w^T x + b = 0$
  - Dual space: $\sum \alpha(i) d(i) x^T x(i) + b = 0$
- Regularization parameter $\gamma$ controls balance between margin and errors.
Fisher Linear Discriminant Analysis

- Based on first and second order statistics of training data. Let $m_{x+}$ ($m_{x-}$) be sample mean of positive (negative) inputs. Let $\Lambda_{x+}$ ($\Lambda_{x-}$) be sample covariance of positive (negative) inputs.
- Project data down to 1 dimension using weight $w$.
- Goal of Fisher LDA is to find $w$ such that $y = \langle w, x \rangle$ and
  - Difference in output means is maximized
    $$|m_{Y+} - m_{Y-}| = |\langle w, m_{x+} - m_{x-} \rangle|$$
  - Minimize within class output covariance
    $$\sigma_{Y+}^2 + \sigma_{Y-}^2$$
Fisher LDA continued

- Define $S_B = (m_{x+} - m_{x-}) (m_{x+} - m_{x-})^T$ as the between class covariance and $S_W = \Lambda_{x+} + \Lambda_{x-}$.

- Fisher LDA can be expressed as finding $w$ to maximize $J(w) = w^T S_B w / w^T S_w w$ (Rayleigh quotient).

- Taking derivative of $J(w)$ with respect to $w$ and setting to zero we get the generalized eigenvalue problem with $S_B w = \lambda S_w w$.

- Solution given by $w = S_w^{-1} (m_{x+} - m_{x-})$ assuming $S_w$ is nonsingular.
Fisher LDA comments

- Fisher LDA projects data down to one dimension by giving optimal weight, w. Threshold value b can be found to give a discriminant function.
- Fisher LDA can also be formulated as a Linear SVM with a quadratic error cost and equality constraints. This gives the Least Squares SVM and adds an additional regularization parameter.
- For Gaussian data with equal covariance matrices and different means, Fisher’s LDA converges to the optimal linear detector.
Implementing Fisher LDA

- $X_1$ is set of positive $m_1$ data and $X_2$ is set of negative $m_2$ data with $m = m_1 + m_2$. Each data item represents one row of matrix.
- Compute first and second order statistics: $m^+ = \text{mean}(X_1)$, $m^- = \text{mean}(X_2)$, $c^+ = \text{cov}(X_1)$, $c^- = \text{cov}(X_2)$. $\text{cov} = (m_1 c^+ + m_2 c^-)/m$;
- $w = (\text{cov})^{-1} (m^+ - m^-)^T$; $b = - (m_1 m^+ + m_2 m^-)^T w/m$;
- Can normalize $w$ and $b$ like SVM so that $m^w + b = 1$. 
Least Squares Algorithm

- Let \((x(k), d(k)), 1 \leq k \leq m\) then LS algorithm finds weight \(w\) such that squared error is minimized. Let \(e(k) = d(k) - w^T x(k)\), then cost function for LS algorithm given by \(J(w) = .5 \sum_k e(k)^2\)

- In matrix form can represent
  \[ J(w) = .5 \|d - Xw\|^2 = .5\|d\|^2 - d^T Xw + .5w^T X^T Xw \]
  where \(d\) is vector of desired outputs and \(X\) contains inputs arranged in rows.
Least Squares Solution

- Let $X$ be the data matrix, $d$ the desired output, and $w$ the weight vector.
- Previously we showed that
  \[ J(w) = 0.5 \|d-Xw\|^2 = 0.5\|d\|^2 - d^T Xw + 0.5w^T X^T Xw \]
  where $d$ is vector of desired outputs and $X$ contains inputs arranged in rows.
- LS solution given by $X^T X w^* = X^T d$ (normal equation) with $w^* = X^+ d$. If $X^T X$ is of full rank then $X^+ = (X^T X)^{-1} X^T$.
- Output $y = Xw^*$ and error $e = d - y$.
- Desired output often of form $d = Xw^* + v$. 
LS Solution Comments

- \( y = Xw + b_1 \), nonzero threshold, solve
  \[ X^T Xw - X^T d - X^T 1b = 0, \quad bm + d^T 1 = w^T X^T 1 \]

- For LS classification positive examples have target value of \( d=1 \) and negative examples have target value of \( d=-1 \).

- Least square solution is same as Fisher discriminant analysis when positive examples have target value \( \frac{m}{m_1} \) and negative examples have target value \( -\frac{m}{m_2} \).

- Can also add regularization: \( J(w) = \frac{1}{2} ||w||^2 + \frac{1}{2} C ||e||^2 \)