Linear SVM Classifier with slack variables (hinge loss function)

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

$$\min_{(w, \xi)} \frac{1}{2}||w||^2 + \gamma \sum \xi(i)$$

subject to $\xi(i) \geq 0 \forall i$, $d(i)(w^T x(i) + b) \geq 1 - \xi(i)$, $\forall i$, and $\gamma > 0$.

In dual space

$$\max W(\alpha) = \sum \alpha(i) - \frac{1}{2} \sum \alpha(i) \alpha(j) d(i) d(j) x(i)^T x(j)$$

subject to $\gamma \geq \alpha(i) \geq 0$, and $\sum \alpha(i) d(i) = 0$.

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

- X1 is set of positive m1 data and X2 is set of negative m2 data with m = m1 + m2. Each data item represents one row of matrix.
- Compute first and second order statistics: m+ = mean(X1), m- = mean(X2), c+ = cov(X1), c- = cov(X2).
  \[ \text{cov} = \frac{(m1 \ c+ + m2 \ c-)}{m}; \]
- \( w = (\text{cov})^{-1} (m+ - m-)^T; b = - (m1 \ m+ + m2 \ m-)^T w/m; \)
- Can normalize w and b like SVM so that m+w + b=1.
Least Squares SVM with nonzero means and regularization

Consider changing SVM to LS SVM by making following modifications:

$$\min_{(w,e)} \frac{1}{2}||w||^2 + \frac{1}{2}C \Sigma e(i)^2$$

subject to $d(i) - (w^T x(i) + b) = e(i)$, $\forall i$, and $C > 0$. Note that $e(i)$ is error term.

Key differences with between SVM and LS SVM:

- For classification problems hinge loss function replaced by quadratic error cost.
- Inequality constraint replaced by equality constraint.
LS SVM solution

- Let $X$ be matrix of training inputs with each $i$th example in row $i$. $X$ is $m$ by $n$. $d$ and $e$ are $m$ vectors and $w$ is an $n$ vector.

- $J(w) = \frac{1}{2}||w||^2 + \frac{1}{2}C||e||^2 = \frac{1}{2}||w||^2 + \frac{1}{2}C||d-Xw-b||^2$

- To find min $J(w)$ take partial derivatives wrt $w$ and $b$ and set to 0.

  1. $(X^TX + I/C) w = X^Td -- X^T1b$

  2. $bm = d^T1 - 1^TXw$

- Substitute $m_x = X^T1/m$, $m_d = d^T1/m$ and (2) into (1) yielding

  - $(X^TX - mm_xm_x^T + I/C) w = X^Td - mm_xm_d$
LS SVM solution comments

- For binary classification, $d$ is -1 or 1.
- Equivalent to Fisher LDA when $C$ is large and $d$ is $-m/m_2$ or $m/m_1$.
- For regression problem, it is sometimes referred to as ridge regression $(C + l/Cm) \hat{w} = Q$ where $C$ is the sample covariance matrix and $Q$ is the cross-covariance vector (approx. to Wiener solution).
- Can also solve problem in dual observation (kernel) space.
- LS SVM involves solving a set of linear equations, however solution not sparse in number of training examples.
Finding Dual Solution

Introduce Lagrange multipliers

\[
L(w,b,e,\alpha) = \frac{1}{2}||w||^2 + \frac{1}{2}C \sum e(i)^2 - \sum \alpha(i) \left( d(i) - (w^T \Phi(x(i)) + b) - e(i) \right)
\]

where \( \alpha(i) \geq 0 \).
KKT Conditions

Again take partial derivatives and set to 0.

\[ \frac{\partial L(w,b,e,\alpha)}{\partial w} = 0, \quad \frac{\partial L(w,b,e,\alpha)}{\partial b} = 0, \]
\[ \frac{\partial L(w,b,e,\alpha)}{\partial \alpha} = \frac{\partial L(w,b,e,\alpha)}{\partial e} = 0. \]

We therefore have that

\[ w = \sum \alpha(i)x(i) \]
\[ \sum \alpha(i) = 0 \]
\[ \alpha(i) = C e(i), \quad 1 \leq i \leq m \]
\[ d(i) - (w^T x(i) + b) - e(i)) = 0, \quad 1 \leq i \leq m \]
Dual Solution to LS SVM

Let $\alpha$ be vector of Lagrange multipliers and $d$ be vector of outputs then solution has following form:

$$
\begin{bmatrix}
0 & 1^T \\
1 & \text{K+I/C}
\end{bmatrix}
\begin{bmatrix}
b \\
\alpha
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
d
\end{bmatrix}
$$

where $K(x,z) = x^T z$ and denotes $l$ vector of 1s.

$$f(x) = \sum \alpha(i) K(x^T x(i)) + b$$
Comments about LS SVM

- Solution to LS SVM depends on $n$, dimensionality of input space $x$ in primal space and $m$, number of training samples in dual space.

- Both solutions involve solving a set of linear equations. Work in space that has lower dimension.

- Adaptive on-line solutions can now be implemented.

- Algorithm easily constructed for pattern classification problems.

- In dual space, all input training examples are support vectors as Lagrange multipliers, $\alpha$ are proportional to error, $e$. 
Capabilities of Linear Threshold Functions

- Discussed three learning algorithms for linear threshold functions (LTF): PLA, SVM, LS SVM (FLDA)
- How can we describe capabilities of LTF? Given m points, how many dichotomies can homogenous LTF (HLTF) (zero threshold) realize?
- General position (GP): m points in $\mathbb{R}^n$ in GP if any subset of $k \leq \min(m, n)$ points are linearly independent.
Function Counting Theorem

Given $m$ points in $\mathbb{R}^n$ in GP there are $C(m,n)$ dichotomies that can be realized where

$$C(m,n) = 2 \sum_{k=0}^{\min(m-1,n-1)} \binom{m-1}{k}$$
FCT Proof

- \( C(m+1,n) = C(m,n) + C(m,n-1) \)
  - Given \( m \) points, add a point \( x^* \) in GP. Construct a hyperplane by projecting into null space of \( x^* \). For any dichotomy, \( x^* \) will either be ambiguous or not. Number of ambiguous points is \( C(m,n-1) \)

- Induction proof:
  - Base step: \( C(m,1) = C(1,n) = 2 \)
  - Induction step
Graphical representation of FCT proof

\[ C(m+1,n) = C(m,n) + C(m,n-1) \]
HLTF capacity is n, LTF capacity is n+1. If points are not in GP capacity is less. Random capacity of HLTF is 2n. Higher capacity achieved by nonlinear threshold functions with capacity dependent on number of inputs. LTF can only realize a limited number of Boolean functions.
3) Regression

- Function output $y$ is real valued.
- Learning algorithm takes labeled training examples and updates parameters $w$ so that some error criterion is minimized.
- Regression problems have different degrees of difficulty requiring appropriate function to perform regression.
- Temporal and spatial learning
Linear Units

A. Preliminaries

\[ y = s = w^T x \]
Model Assumptions and Parameters

- Training examples \((x(k),d(k))\) drawn randomly, second order zero mean sequences.
- Parameters
  - Inputs: \(x(k) \in \mathbb{R}^n\)
  - Weights: \(w(k) \in \mathbb{R}^n\)
  - Outputs: \(y(k) = w(k)^T x(k)\)
  - Desired outputs: \(d(k)\)
  - Error: \(e(k) = d(k) - y(k)\)
- Error criterion (MSE)
  \[
  \min J(w) = E \left[ .5(e(k))^2 \right]
  \]
Define $P = E(x(k)d(k))$ and $R = E(x(k)x(k)^T)$.

$$J(w) = 0.5 \ E[(d(k)-y(k))^2]$$

$$= 0.5 E(d(k)^2) - E(x(k)d(k))^T w + w^T E(x(k)x(k)^T) w$$

$$= 0.5 E[d(k)^2] - P^T w + 0.5 w^T R w$$

Note $J(w)$ is a quadratic function of $w$. To minimize $J(w)$ find gradient, $\nabla J(w)$ and set to 0.

$$\nabla J(w) = -P + Rw = 0$$

$Rw = P$ (Wiener solution)

If $R$ is nonsingular, then $w = R^{-1}P$.

Resulting MSE $= 0.5E[d(k)^2] - 0.5P^T R^{-1}P$
Gradient based iterative algorithms

- Steepest descent algorithm (move in direction of negative gradient)
  \[ w(k+1) = w(k) - \mu \nabla J(w(k)) = w(k) + \mu (P - Rw(k)) \]

- Least mean square algorithm (approximate gradient from training example)
  \[ \nabla J(w(k)) = -e(k)x(k) \]
  \[ w(k+1) = w(k) + \mu e(k)x(k) \]
Steepest Descent Convergence

- $w(k+1) = w(k) + \mu (P-Rw(k))$; Let $w^*$ be solution.
  Center weight vector $v = w - w^*$
- $v(k+1) = v(k) - \mu (Rv(k))$; Assume $R$ is nonsingular.
  Decorrelate weight vector $u = Q^{-1}v$ where $R = Q\Lambda Q^{-1}$ is the transformation that diagonalizes $R$.
- $u(k+1) = (I - \mu \Lambda)u(k)$, $u(k) = (I - \mu \Lambda)^k u(0)$.
  Conditions for convergence $0 < \mu < 2/\lambda_{\text{max}}$. 
Step Size $\mu$

- $\mu$ too large
- $\mu$ too small
Rate of Convergence

- Rate of convergence depends on eigenvalues, $\lambda_i$ as convergence rate for this eigenvalue is $(1 - \mu \lambda_i)$. Key eigenvalues are $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$.
- Fastest rate of convergence achieved when setting $\mu = 2 / (\lambda_{\text{min}} + \lambda_{\text{max}})$. This results in smallest and largest eigenvalue having same convergence rate.
- Convergence of SD depends on condition number of matrix $\lambda_{\text{max}} / \lambda_{\text{min}}$. 
Energy Function

- Energy Function:
  \[ J(w) = 0.5\sigma_d^2 - P^T w + 0.5w^T R w \]

For optimal weight \( R w^* = P \) and

\[ J_{\text{min}} = J(w^*) = 0.5\sigma_d^2 - 0.5 P^T w^* \]

- SD energy function behavior

\[ J(w(k)) = J_{\text{min}} + 0.5 (w(k)-w^*)^T R (w(k)-w^*) \]
\[ = J_{\text{min}} + 0.5u(k)^T \Lambda u(k) \]
\[ = J_{\text{min}} + 0.5\sum_i (1 - \mu \lambda_i)^{2k} u_i(0)^2 \]