Formulation with slack variables

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 $w = \sum \alpha(i) d(i) x(i)$.
QP software for SVM

- Matlab (easy to use, choose primal or dual space, slow): quadprog()
  - Primal space (w,b, ξ+, ξ-)
  - Dual space (α)
- 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 + pts. Mean = (.5,.5)
20 - pts. Mean = -(.5,.5)
Example continued

Primal Space (matlab)

```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)

% Example code for dual space in MATLAB

xn = x.* (d*[1 1]);
k= xn*xn';
gamma = 1;
f= -ones(40,1);
Aeq = d';
beq = 0
A = zeros(1,40);
b = 0;
lb = [zeros(40,1)];
ub = [gamma*ones(40,1)];
[alpha,fvala] = quadprog(k,f,A,b,Aeq,beq,lb,ub);
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.
SMO Algorithm

Sequential Minimization Optimization breaks up QP program into small subproblems that are solved analytically.

- SMO solves dual QP SVM problem by examining points that violate KKT conditions.
- Algorithm converges and consists of:
  - Search for 2 points that violate KKT conditions.
  - Solve QP program for 2 points.
  - Calculate threshold value \( b \).
- Continue until all points satisfy KKT conditions.
- Convergence time depends on difficulty of classification problem and kernel functions used.
SVM Light

- Decomposition method: partition data into small working set and fixed set. Only learn Lagrangian multipliers associated with training data belonging to working set.
- Software available at Thorsten Joachim’s website at Cornell University.
- Data represented to take advantage of possible sparseness, nonnumeric training data.
References

- Kernel website: http://www.kernel-machines.org/
Optimal Bayesian Classification

- Probability models known: Prior distributions, conditional distributions
- Costs associated with making decision
- Minimize expected cost function
- Decision based on sufficient statistic (Likelihood function)

Also known as Detection Theory, Decision Theory
Decision Model

- Bayesian model: Priors, costs, distributions known
- Learning model: Training data, distributions unknown
Bayesian Detection

- Two hypothesis detection problem: X inputs, D classes.
  - $D = -1$, null hypothesis, $f_{X|D}(x|-1)$.
  - $D = 1$, alternate hypothesis, $f_{X|D}(x|1)$.

- Bayesian model
  - Priors known $P(D=-1)=p$, $P(D=1)=1-p$
  - Costs: $C(i,j)$ cost of deciding $Y=i$ given $D=j$

- Objective: partition input space into two sets to minimize average cost
Bayes risk

Proposition: Bayes average risk is minimized by following decision rule

\[ f(x) = \text{sign} \left( L(x) - L_0 \right) \]

where \( L(x) = \frac{f_{X|D}(x|1)}{f_{X|D}(x|-1)} \) is the likelihood ratio and threshold value given by

\[ L_0 = \frac{p/(1-p)}{(C(1,-1)-C(-1,-1))/(C(-1,1)-C(1,1))} \]

Comment: Can also work with log likelihood function \( l(x) = \log(L(x)) \).
Minimum Error Probability

- Costs: $C(-1,-1) = C(1,1) = 0$ and $C(-1,1) = C(1,-1) = 1$.
- Minimum Error probability is equivalent to MAP decision rule
  \[
  \max_i P(D=i|x) = \max_i f_{X|D}(x|i) \frac{P(D=i)}{f_X(x)}
  \]
  from Bayes relationship

- For Gaussian RVs with different means and the same covariance matrix decision rule is given by a linear threshold detector. (Sufficient statistic $t(x) = s^T \Lambda^{-1} x$)
- For Gaussian RVs with different means and different covariance matrix decision rule is given by a quadratic threshold detector.