Lecture 17: Modern Classification (II) - Support Vector Machines

Hao Helen Zhang

Outlines

- Linear SVM for Non-separable Problems
  - Two classes overlap
- Nonlinear SVM
  - Data can not be separated well by linear SVM
  - Data may be separated well by linearly SVM after some nonlinear transformation
  - Linear SVM in the new input space (after transformation) implies nonlinear SVM in the original input space

Optimal Separating Hyperplane

The linear SVM for perfectly separable cases:

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|^2$$

$$\text{subject to} \quad y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq 1, \quad \text{for} \quad i = 1, ..., n$$

For non-separable data,

- Such a pair $(\beta_0, \boldsymbol{\beta})$ does not exist!
- Why?
- How to generalize the linear SVM here?

Soft Margin Hyperplane

Relaxed constraints by introducing positive variables $\xi_i \geq 0$ for each i.

$$\begin{aligned} \mathbf{x}^T\boldsymbol{\beta} + \beta_0 &\geq 1 - \xi_i \quad \text{for} \quad y_i = 1 \\ \mathbf{x}^T\boldsymbol{\beta} + \beta_0 &\leq -1 + \xi_i \quad \text{for} \quad y_i = -1 \end{aligned}$$

Basic ideas:

- Whenever the constraint $\mathbf{x}^T\boldsymbol{\beta} + \beta_0 \geq 1$ is violated for some point in $+1$ class, we lend it a positive constant ξ_i to make the constraint hold.
- Similar for the violations in -1 class.

The above two inequalities can be summarized as

$$y_i(\mathbf{x}^T\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, \quad \forall i = 1, \cdots, n.$$

Optimization Problem for SVM

Linear SVM for non-separable problems:

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}} \quad & \tfrac{1}{2}||\boldsymbol{\beta}||^2 + \gamma(\textstyle\sum_{i=1}^n \xi_i) \\ \text{subject to} \quad & y_i(\mathbf{x}^T\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, \quad i = 1, ..., n. \\ & \xi_i \geq 0, \quad i = 1, ..., n, \end{aligned}$$

for some $\gamma > 0$, a penalty to errors.

- If an error occurs at the training point $(\mathbf{x}_i, y_i)$, then $\xi_i > 1$.
- The quantity $\sum_i \xi_i$ is an upper bound on the number of training errors.

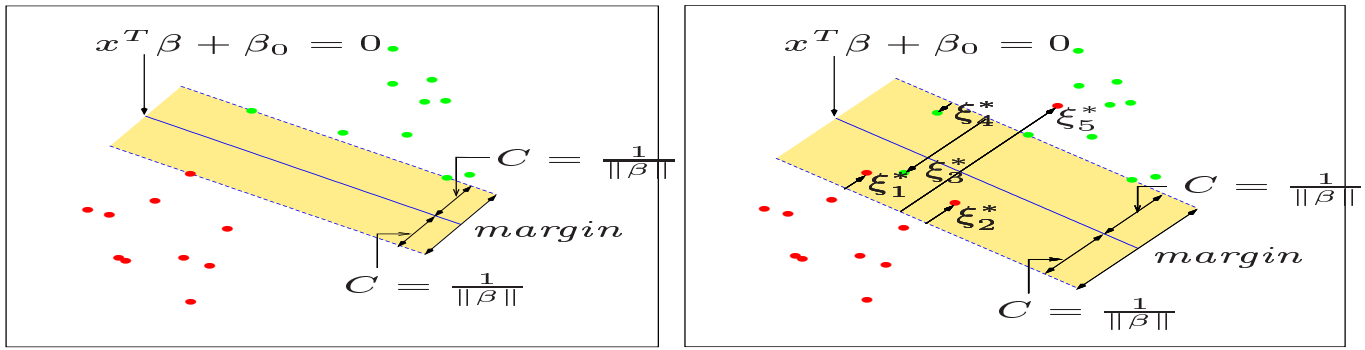


Figure 12.1: *Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width* $2C = 2/\|\beta\|$. *The right panel shows the nonseparable (overlap) case. The points labeled* ξ_j^* *are on the wrong side of their margin by an amount* $\xi_j^* = C\xi_j$; *points on the correct side have* $\xi_j^* = 0$. *The margin is maximized subject to a total budget* $\sum \xi_i \leq$ *constant. Hence* $\sum \xi_j^*$ *is the total distance of points on the wrong side of their margin.*

Regularization/Tuning Parameter

We say γ the *regularization*, or tuning, or cost parameter.

- γ balances the training error (bound) and the margin width
- In the separable case, $\gamma = \infty$. (why?)

In literature, C is sometimes used as the tuning parameter

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}} \quad & \tfrac{1}{2}||\boldsymbol{\beta}||^2 + C(\textstyle\sum_{i=1}^n \xi_i) \\ \text{subject to} \quad & y_i(\mathbf{x}^T\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, \quad i = 1, ..., n. \\ & \xi_i \geq 0, \quad i = 1, ..., n. \end{aligned}$$

The SVM optimization is a QP.

Quadratic Programming (QP)

Introducing the Lagrange multipliers $\alpha_i \geq 0, \mu_i \geq 0$ for each constraint, we obtain the Lagrange function

$$\begin{aligned} L(\boldsymbol{\beta}, \beta_0, \boldsymbol{\alpha}, \xi, \mu) \quad \equiv \quad & \frac{1}{2}||\boldsymbol{\beta}||^2 + \gamma \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \mu_i \xi_i \\ & - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - 1 + \xi_i] \end{aligned}$$

Optimization problems:

- minimize L with respect to *primal* variables $\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}$
- maximize L with respect to *dual* variables $\boldsymbol{\alpha}, \boldsymbol{\mu}$.

Stationary Points

Using the KKT theory, the optimal solution should satisfy the following equations

$$\frac{\partial}{\partial \beta_0} L = 0, \quad \frac{\partial}{\partial \boldsymbol{\beta}} L = 0, \quad \frac{\partial}{\partial \xi_i} L = 0.$$

We get

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \alpha_i y_i \\ \boldsymbol{\beta} &= \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \alpha_i &= \gamma - \mu_i, \quad i = 1, ..., n \end{aligned}$$

KKT Conditions for Soft-margin SVM

- Stationary conditions:

$$0 = \sum_{i=1}^{n} \alpha_i y_i, \quad \boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \quad \alpha_i = \gamma - \mu_i, \quad i = 1, ..., n.$$

- Primal feasible:

$$y_i(\mathbf{x}^T \boldsymbol{\beta} + \beta_0) - (1 - \xi_i) \geq 0, \quad \xi_i \geq 0, \quad i = 1, ..., n.$$

- Dual feasible:

$$\alpha_i \geq 0, \quad \mu_i \geq 0, \quad i = 1, ..., n.$$

- Complementary slackness:

$$0 = \alpha_i[y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - 1 + \xi_i], \quad 0 = \mu_i \xi_i, \quad i = 1, ..., n.$$

They together **uniquely** characterize the solution to primal/dual problem.

Solution Properties

- If $y_i(\beta^T \mathbf{x}_i + \beta_0) > 1$ (points outside the margin)
  - points are correctly classified
  - From the complementary slackness condition, we have $\alpha_i = 0$. It implies that $\mu_i = \gamma$ and $\xi_i = 0$.
- If $y_i(\beta^T \mathbf{x}_i + \beta_0) = 1$ (points on the margin), then $\xi_i = 0$.
  - If $\alpha_i = 0$, then $\mu_i = \gamma$ and $\xi_i = 0$.
  - If $\alpha_i > 0$, then $\xi_i = 0$.
- If $y_i(\beta^T \mathbf{x}_i + \beta_0) < 1$ (points within the margin), then $\xi_i > 0$.
  - From the complementary slackness condition, we have $\alpha_i = \gamma$. (Why?)
  - Include two types of points: correctly classified (but in the margin band; misclassified points)

Support Vectors

The set of support vectors: $SV = \{i | \hat{\alpha}_i > 0, \quad i = 1..., n\}$

$$\hat{\boldsymbol{\beta}} = \sum_{i \in SV} \hat{\alpha}_i y_i \mathbf{x}_i$$

Three types of SVs:

- correctly classified points on the margin, $\xi_i = 0$ and $0 < \alpha_i < \gamma$
- Correctly classified points within the margin, $0 < \xi_i < 1$ and $\alpha_i = \gamma$.
- Misclassified points with $\xi_i > 1$, hence $\alpha_i = \gamma$.

Dual Problem

By substituting both into L_P , we get

$$G(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

subject to

$$0 \leq \alpha_i \leq \gamma, \quad i = 1, ..., n;$$
$$\sum_i \alpha_i y_i = 0.$$

Maximizing the dual is a simpler convex QP than the primal.

Obtain SVM Primal Solutions

- To solve $\boldsymbol{\beta}$:
  - Define the dual solution as $\boldsymbol{\alpha}^*$, then

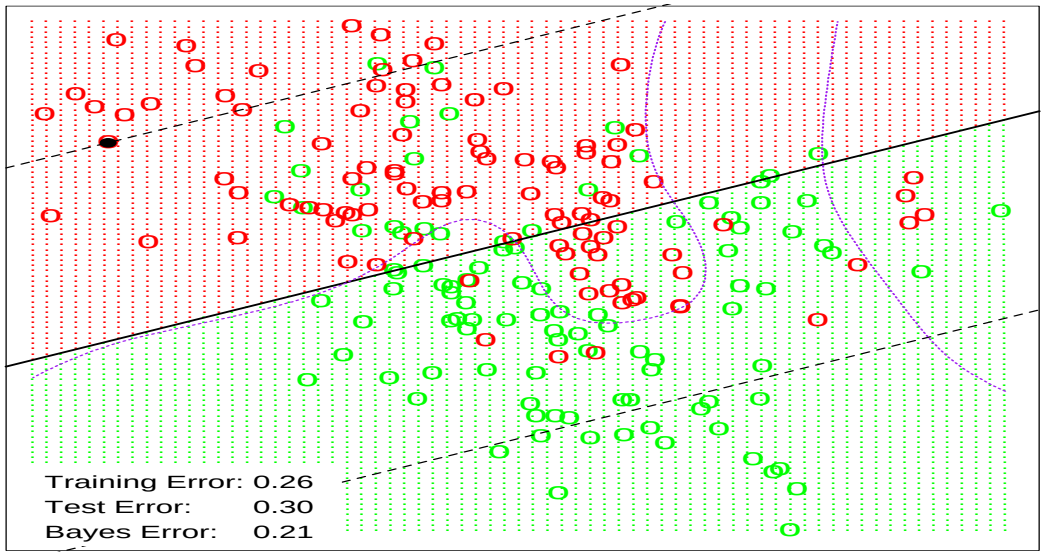
$$\boldsymbol{\beta}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i.$$

- To solve β_0
  - Use any point satisfying $\alpha_i > 0, \xi_i = 0$.
  - Typically use an average of all the solutions for numerical stability.

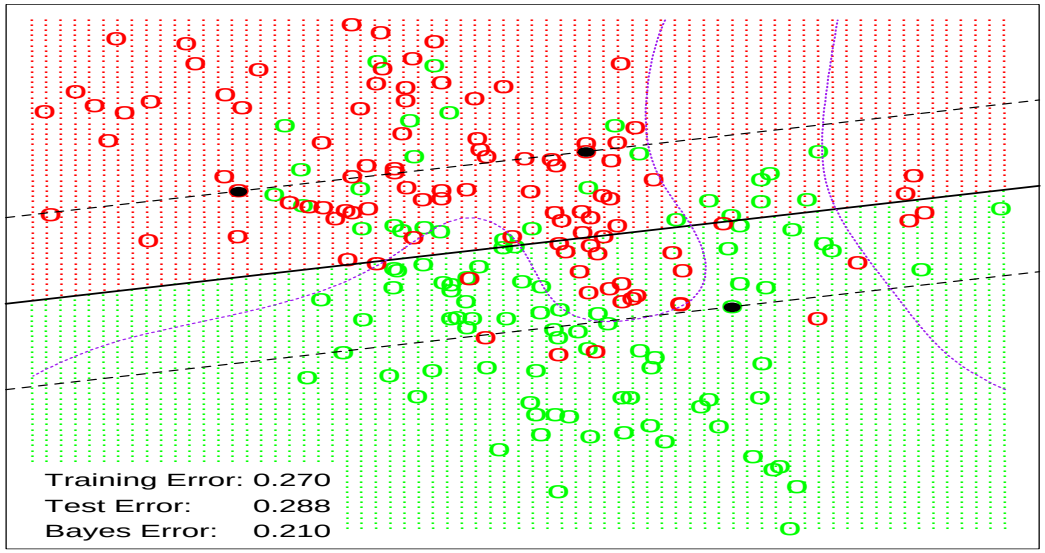
Tuning Parameter γ

- large γ puts more weight on misclassification rate than margin width
  - discourage any positive ξ_i
  - may lead to an overfit wiggly boundary in the original space
- small γ puts more weight on margin width than misclassification rate
  - encourage small value of $\|\beta\|$
  - may lead to smoother boundary

Tuning procedures: tuning set, cross-validation; leave-one-out cross validation



$\gamma = 0.01$



$\gamma = 10000$

Figure 12.2: *The linear support vector boundary for the mixture data example with two overlapping classes, for two different values of γ. The broken lines indicate the margins, where $f(x) = \pm 1$. The support points ($\alpha_i > 0$) are all the points on the wrong side of their margin. The black solid dots are those support points falling exactly on the margin ($\xi_i = 0$, $\alpha_i > 0$). In the upper panel 62% of the observations are support points, while in the lower panel 85% are.*

Nonlinear Support Vector Machines

Allow more general decision surfaces

- Nonlinearly map data into some other inner product space

$$\Phi : R^d \to \mathcal{F}$$

- Select basis functions of $\mathcal{F}$: $h_m(\mathbf{x}), m = 1, ..., M$.
- Fit the SVM classifier using features

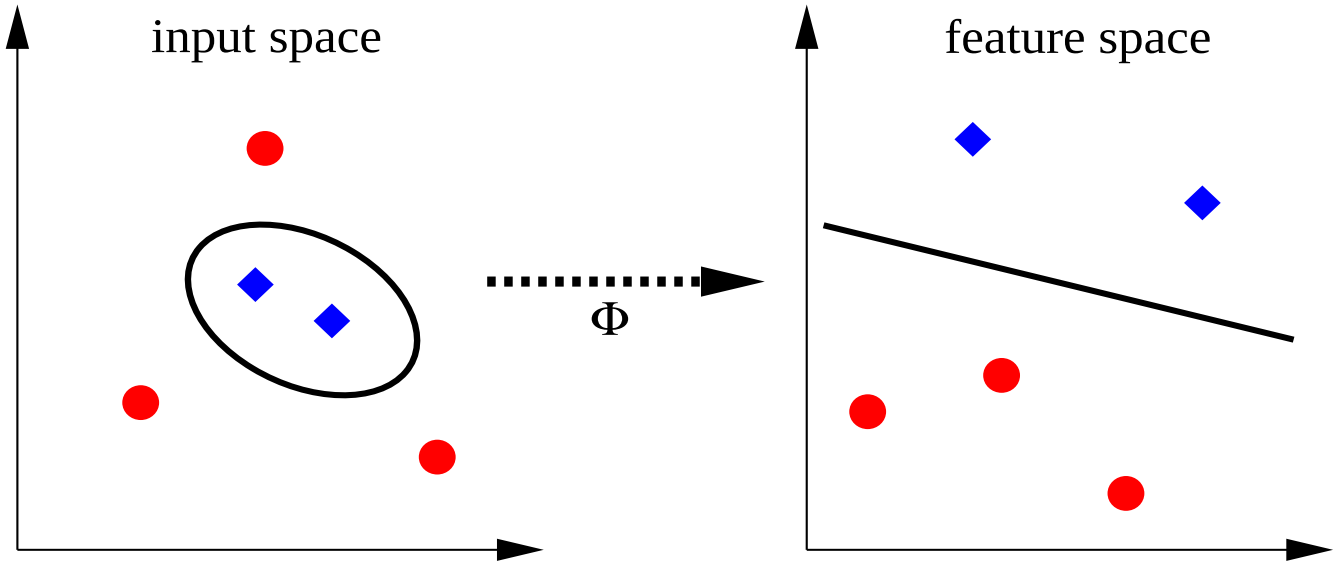
$$h(\mathbf{x}_i) = (h_1(\mathbf{x}_i), h_2(\mathbf{x}_i), ..., h_M(\mathbf{x}_i))$$

- Produce the nonlinear function $\hat{f}(\mathbf{x}) = h(\mathbf{x})^T \hat{\boldsymbol{\beta}} + \hat{\beta}_0$.
- The final classifier: $\hat{y}(\mathbf{x}) = \text{sign}(\hat{f}(\mathbf{x}))$.

The dimension of the enlarged space can have very large, infinite dimensions.

- The computation can be prohibitive.

Support Vector Classifiers



[6]

B. Schölkopf, Canberra, February 2002

Dual Problems

The Lagrange dual function

$$G(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j < h(\mathbf{x}_i), h(\mathbf{x}_j) >$$

subject to

$$0 \leq \alpha_i \leq \gamma, \quad i = 1, ..., n; \quad \sum_i \alpha_i y_i = 0.$$

The solution function

$$f(\mathbf{x}) = h(\mathbf{x})^T \boldsymbol{\beta} + \beta_0 = \sum_{i=1}^{n} \alpha_i y_i < h(\mathbf{x}), h(\mathbf{x}_i) > + \beta_0.$$

Given α_i, the intercept β_0 is determined by solving $f(\mathbf{x}_i) = 0$ for any $\mathbf{x}_i$ for which $0 < \alpha_i < \gamma$.

Kernel Function

The prediction function only relies on $< h(\mathbf{x}), h(\mathbf{x}_i) >$.
We define this inner product at *kernel* function

$$K(\mathbf{x}, \mathbf{x}') =< h(\mathbf{x}), h(\mathbf{x}') >$$

- K symmetric positive (semi-) definite function
- Various kernels are used in literature:

dth Degree polynomial: $K(\mathbf{x}, \mathbf{x}') = (1+ < \mathbf{x}, \mathbf{x}' >)^d$

Radial basis: $K(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/\sigma)$

Neural network: $K(\mathbf{x}, \mathbf{x}') = \tanh(\kappa_1 < \mathbf{x}, \mathbf{x}' > +\kappa_2)$

Polynomial Kernel

For $d = 2$, the kernel function value (x_1, x_2) and (x_1', x_2')

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &= (1+ < \mathbf{x}, \mathbf{x}' >)^2 \\ &= (1 + x_1 x_1' + x_2' x_2')^2 \\ &= 1 + 2x_1 x_1' + 2x_2' x_2' + (x_1 x_1')^2 + (x_2' x_2')^2 + 2x_1 x_1' x_2' x_2' \end{aligned}$$

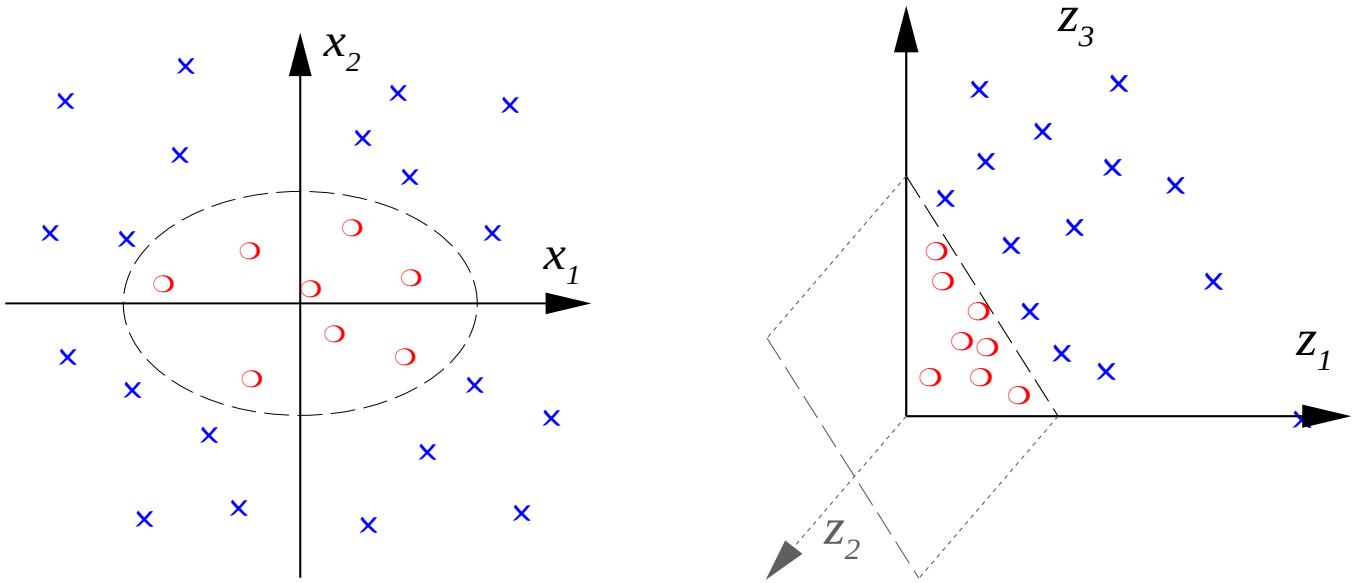
Then $M = 6$, with

$$\begin{array}{lll} h_1(\mathbf{x}) = 1 & h_2(\mathbf{x}) = \sqrt{2}x_1 & h_3(\mathbf{x}) = \sqrt{2}x_2 \\ h_4(\mathbf{x}) = x_1^2 & h_5(\mathbf{x}) = x_2^2 & h_6(\mathbf{x}) = \sqrt{2}x_1 x_2 \end{array}$$

Check: $K(\mathbf{x}, \mathbf{x}') =< h(\mathbf{x}), h(\mathbf{x}') >$.

Example: All Degree 2 Monomials

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$



B. Schölkopf, Canberra, February 2002

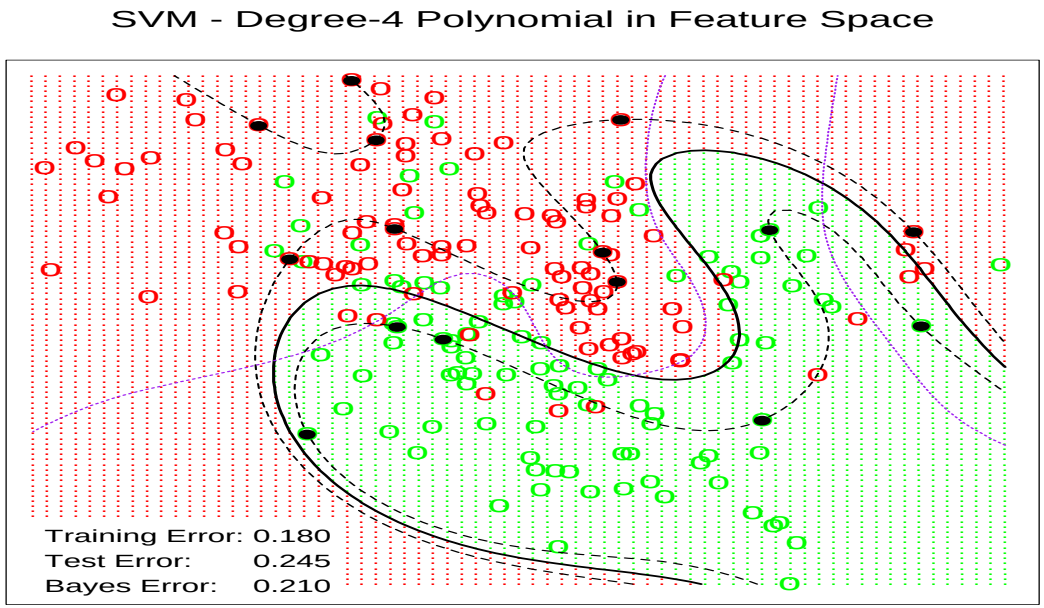
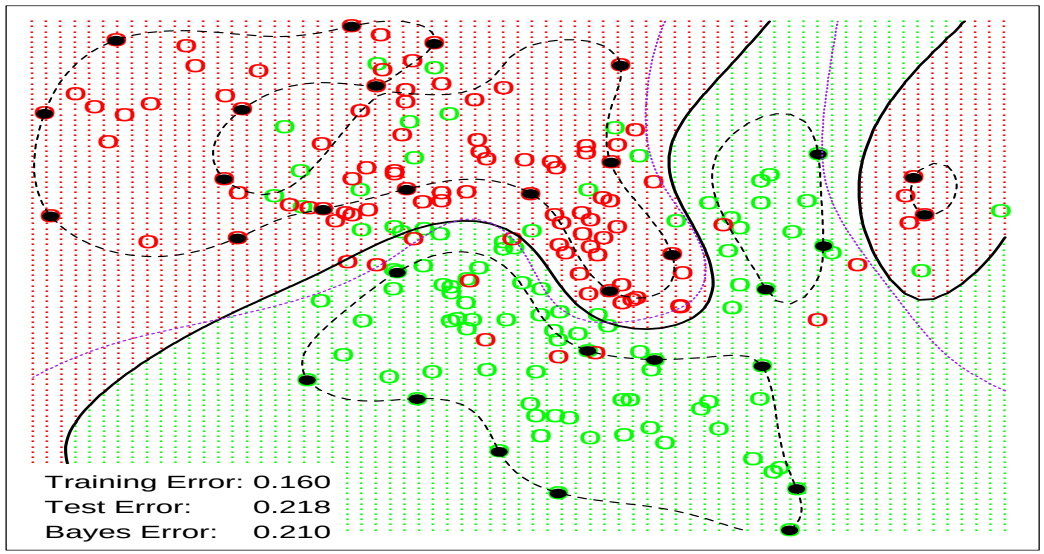


Figure 12.3: *Two nonlinear SVMs for the mixture data. The upper plot uses a 4-th degree polynomial kernel, the lower a radial basis kernel. In each case γ was tuned to approximately achieve the best test error performance, and $\gamma = 1$ worked well in both cases. The radial basis kernel performs the best (close to Bayes optimal), as might be expected given the data arise from mixtures of Gaussians. The broken purple curve in the background is the Bayes decision boundary.*

Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001 Chapter 12

SVM - Radial Kernel in Feature Space



R Functions: SVM

- R package *kernlab*; function *ksvm*.
- R package *e1071*; function *svm*.
- R package *svmpath*: compute the entire regularized solution path.