

MATH 680 Computation Intensive Statistics

October 9, 2024

SVM

Contents

1	Optimal separating hyperplanes	1
2	Overlapping class distributions	6
3	Support Vector Machines and Kernels	13
4	The SVM as A Penalization Method	16
5	Kernel Regression	19

1 Optimal separating hyperplanes

Consider two-class classification problem, where we find a **hyperplane** or affine set L defined by the equation $f(x) = \beta_0 + \beta^\top x = 0$. The classifier is $\text{sign}(x^\top \beta + \beta_0)$. It has some important properties that can help us formula the SVM:

1. For any x_1 and x_2 lying in L

$$\beta^\top(x_1 - x_2) = 0$$

and hence

$$\beta^* = \frac{\beta}{\|\beta\|}$$

is the **unit vector** normal (perpendicular) to the surface of L .

2. For any point x_0 in L , $\beta^\top x_0 = -\beta_0$.

3. The **signed distance** of any point x to L is

$$\begin{aligned}\langle \beta^*, x - x_0 \rangle &= \beta^{*\top}(x - x_0) = \frac{\beta^\top}{\|\beta\|}(x - x_0) = \frac{1}{\|\beta\|}(\beta^\top x + \beta_0) \\ &= \frac{1}{\|f'(x)\|}f(x),\end{aligned}$$

where x_0 is a point in L .

With preparation of those notions, we are ready to introduce the SVM. For simplicity, let us first consider the well-separated case, where there is no overlapping points between two classes, while the non-overlapping case will be discussed later. We hope to find a separating hyperplane (which characterized by β and β_0) by maximizing the distance of points to the decision boundary, those points are all correctly classified.

We observe that the correctly classified point i has the following property,

- when true $y_i = +1$, the predicted $\hat{y}_i = f(x_i) = x_i^\top \beta + \beta_0 > 0$ (half space)
- when true $y_i = -1$, the predicted $\hat{y}_i = f(x_i) = x_i^\top \beta + \beta_0 < 0$ (half space)

For both cases, the **distance** of a correctly classified x_i to the decision surface is given by

$$\frac{y_i(\beta^\top x_i + \beta_0)}{\|\beta\|} > 0$$

We hope to achieve the following two goals:

- **Correctly classification:** we require that all the points (x_i, y_i) for $i = 1, \dots, N$ be correctly classified (later we will relax this requirements by introducing slackness).
- **Maximized margin:** the lower-bound of the respective distances of N points to the hyperplane (i.e. the margin) should be maximized. We define lower-bound as **margin** $M > 0$, which basically is the perpendicular distance from the hyperplane to the closest point x_i .

Thus the problem of maximizing the margin of the separating hyperplane can be formulated as the following optimization problem

$$\begin{aligned} & \max_{(\beta, \beta_0)} M \\ \text{subject to } & \frac{1}{\|\beta\|} y_i(\beta^\top x_i + \beta_0) \geq M, \quad i = 1, \dots, N. \end{aligned} \quad (1)$$

The above problem can be further simplified. Observe that any positively scaled $\beta \rightarrow \kappa\beta$ and $\beta_0 \rightarrow \kappa\beta_0$ would still satisfy the inequality constraint:

$$\frac{1}{\|\beta\|} y_i(\beta^\top x_i + \beta_0) \geq M$$

Therefore we can arbitrarily set $\|\beta\| = 1/M$. Thus (1) can be simplified to

$$\begin{aligned} & \max_{\beta, \beta_0} \frac{1}{\|\beta\|} \\ & \text{subject to } y_i(\beta^\top x_i + \beta_0) \geq 1, \quad i = 1, \dots, N. \end{aligned}$$

which is equivalent to

$$\begin{aligned} & \min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 \\ & \text{subject to } y_i(\beta^\top x_i + \beta_0) \geq 1, \quad i = 1, \dots, N. \end{aligned} \quad (2)$$

This is a convex quadratic optimization problem, in which we minimize a quadratic function subject to a set of linear inequality constraints. The corresponding **Lagrangian** is

$$L(\beta, \beta_0, \alpha) = \frac{1}{2} \|\beta\|^2 + \sum_{i=1}^N \alpha_i [1 - y_i(\beta^\top x_i + \beta_0)], \quad (3)$$

with $\alpha_i \geq 0$. Now we can find the dual problem of (2), which is the maximization problem of the **Lagrange dual function** $L_D(\alpha)$ in terms of the dual variable α , subjecting to the constraint for the dual variable $\alpha_i \geq 0$ for $i = 1, \dots, N$. $L_D(\alpha)$ is defined as follows

$$L_D(\alpha) = \min_{(\beta, \beta_0)} L(\beta, \beta_0, \alpha) \quad (4)$$

The dual problem therefore is

$$\begin{aligned} & \max_{\alpha} L_D(\alpha) = \max_{\alpha} \min_{(\beta, \beta_0)} L(\beta, \beta_0, \alpha) \\ & \text{subject to } \alpha_i \geq 0 \end{aligned} \quad (5)$$

This dual problem can be further simplified. Observing that at the solution of the primal problem (2) $\beta = \beta^*$ and $\beta_0 = \beta_0^*$, the stationarity condition holds $\frac{\partial L(\beta, \beta_0, \alpha)}{\partial \beta_0} \Big|_{\beta=\beta^*} = 0$ and

$$\frac{\partial L(\beta, \beta_0, \alpha)}{\partial \beta} \Big|_{\beta_0=\beta_0^*} = 0, \text{ which yields}$$

$$\frac{\partial L(\beta, \beta_0, \alpha)}{\partial \beta_0} \Big|_{\beta=\beta^*} = 0 \implies \beta^* = \sum_{i=1}^N \alpha_i y_i x_i, \quad (6)$$

$$\frac{\partial L(\beta, \beta_0, \alpha)}{\partial \beta} \Big|_{\beta_0=\beta_0^*} = 0 \implies 0 = \sum_{i=1}^N \alpha_i y_i, \quad (7)$$

and plug these in (4), we obtain a simplified version of $L_D(\alpha)$

$$L_D(\alpha) = \min_{(\beta, \beta_0)} L(\beta, \beta_0, \alpha) = L(\beta^*, \beta_0^*, \alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{k=1}^N \alpha_i \alpha_k y_i y_k x_i^\top x_k$$

Hence the dual problem (5) can be simplified into

$$\begin{aligned} \max_{\alpha} L_D(\alpha) &= \max_{\alpha} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{k=1}^N \alpha_i \alpha_k y_i y_k x_i^\top x_k \\ \text{subject to } \alpha_i &\geq 0 \text{ and } \sum_{i=1}^N \alpha_i y_i = 0. \end{aligned} \quad (8)$$

This is the dual form of the non-overlapping case of the linear SVM.

We can further get insight into the property of the primal and dual solutions. The solution β^* , β_0^* and α^* of the above primal and dual problems must satisfy the Karush-Kuhn-Tucker conditions, which include

- **stationarity:** (6) and (7),
- **dual feasibility:** (8) and

- complementary slackness:

$$\alpha_i^*[y_i(x_i^\top \beta^* + \beta_0^*) - 1] = 0 \quad \forall i, \quad (9)$$

from which we see that

- If $\alpha_i^* > 0$, to satisfy (9), it must be that $y_i(x_i^\top \beta^* + \beta_0^*) = 1$, x_i is on the boundary of the margin.
- If $y_i(x_i^\top \beta^* + \beta_0^*) > 1$, then x_i is not on the boundary of the margin, then it must be that $\alpha_i^* = 0$.

We see that the solution vector β^* and β_0^* is defined in terms of a linear combination of the support point x_i , where corresponding $\alpha_i^* > 0$. To predict new data points using the trained model, we evaluate the sign $\text{sign}(f(x)) = \text{sign}(x^\top \beta^* + \beta_0^*)$. Plug in (6), we get

$$f(x) = \sum_{i=1}^N \alpha_i^* y_i \langle x_i, x \rangle + \beta_0^*. \quad (10)$$

Data points with $\alpha_i > 0$, are called **support vector**; Data points with $\alpha_i = 0$, plays no role in (10).

2 Overlapping class distributions

For the general data set, there is no $x^\top \beta + \beta_0 = 0$ to perfectly separate class “+1” from class “-1”.

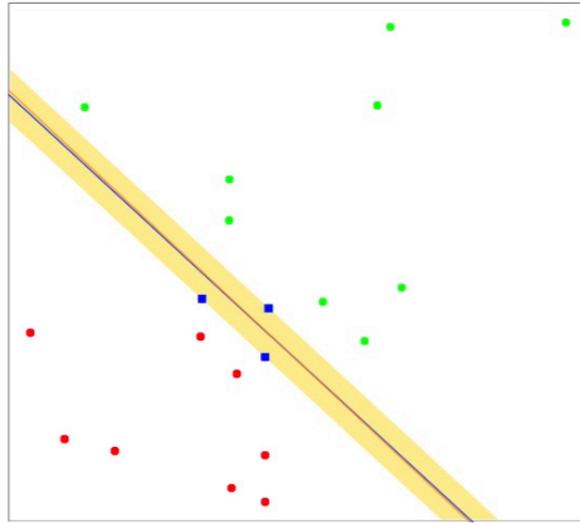


Figure 1: The same data as in Figure ???. The shaded region delineates the maximum margin separating the two classes. There are three support points indicated, which lie on the boundary of the margin, and the optimal separating hyperplane (blue line) bisects the slab. Included in the figure is the boundary found using logistic regression (red line), which is very close to the optimal separating hyperplane.

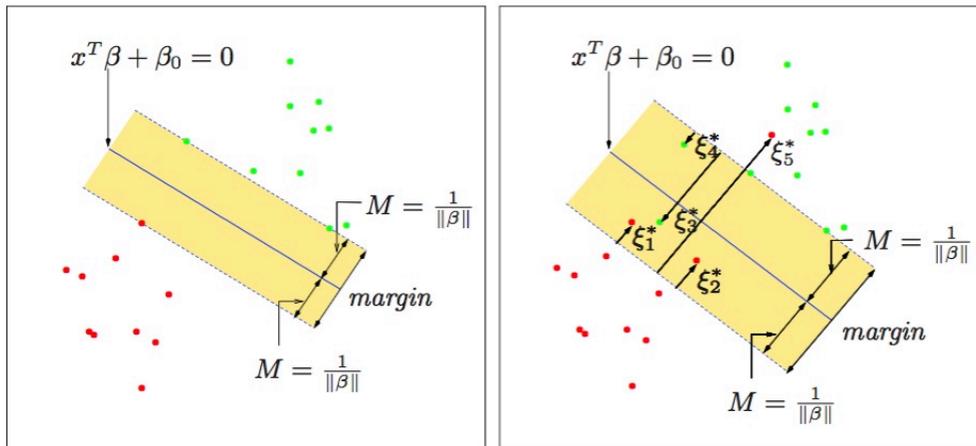


Figure 2: Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width $2M = 2/\|\beta\|$. The right panel shows the nonseparable (overlap) case. The points labeled ξ_j^* are on the wrong side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

The original optimization problem is

$$\begin{aligned} & \max_{(\beta, \beta_0)} M \\ \text{subject to } & \frac{1}{\|\beta\|} y_i (\beta^\top x_i + \beta_0) \geq M, \quad i = 1, \dots, N. \end{aligned} \quad (11)$$

Suppose now that the classes overlap in feature space. We now allow for some points to be on the wrong side of the margin. Define the **slack variables** $\xi = (\xi_1, \xi_2, \dots, \xi_N)$. There is a natural way to modify the constraint in (11):

$$\frac{1}{\|\beta\|} y_i (\beta^\top x_i + \beta_0) \geq M(1 - \xi_i) \quad \forall i, \quad \xi_i \geq 0, \quad \sum_{i=1}^N \xi_i \leq \text{const.}$$

Misclassifications occur when $\xi_i > 1$. Define $M = 1/\|\beta\|$, we get

$$\begin{aligned} & \min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 \\ \text{subject to } & y_i (\beta^\top x_i + \beta_0) \geq 1 - \xi_i \quad \forall i, \quad \xi_i \geq 0, \quad \sum_{i=1}^N \xi_i \leq \text{const.} \end{aligned}$$

The problem is quadratic with linear inequality constraints, hence is convex optimization problem. For convenience, we re-express it as

$$\begin{aligned} & \min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \xi_i \\ \text{subject to } & y_i (\beta^\top x_i + \beta_0) \geq 1 - \xi_i \quad \forall i, \quad \xi_i \geq 0 \end{aligned} \quad (12)$$

We know that

$$\begin{aligned}
& y_i(x_i^\top \beta + \beta_0) + \xi_i \geq 1, \forall i \\
& \xi_i \geq 0 \\
& \iff (1 - \xi_i - y_i(x_i^\top \beta + \beta_0)) \leq 0, \forall i, \\
& -\xi_i \leq 0.
\end{aligned}$$

We obtain the **Lagrangian** as

$$L(\beta, \beta_0, \xi, \mu, \alpha) = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \xi_i + \sum_{i=1}^N \alpha_i [(1 - \xi_i) - y_i(x_i^\top \beta + \beta_0)] - \sum_{i=1}^N \mu_i \xi_i. \tag{13}$$

$$\frac{\partial L}{\partial \beta} = \beta - \sum_{i=1}^N \alpha_i y_i x_i = 0 \tag{14}$$

$$\frac{\partial L}{\partial \beta_0} = - \sum_{i=1}^N \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0$$

$$\iff \begin{cases} \alpha_i = C - \mu_i \\ \sum_{i=1}^N \alpha_i y_i = 0 \\ \sum_{i=1}^N \alpha_i y_i x_i = \beta \end{cases} \tag{15}$$

and $\alpha_i \geq 0, \mu_i \geq 0, \xi_i \geq 0 \forall i$. Plug in (15) into (13) we get the **Lagrange dual function**

as $L_D(\mu, \alpha) = \min_{\beta, \beta_0, \xi} L(\beta, \beta_0, \xi, \mu, \alpha)$.

$$\begin{aligned}
L_D(\mu, \alpha) &= \frac{1}{2} \left\| \sum_{i=1}^N \alpha_i y_i x_i \right\|_2^2 + \sum_{i=1}^N \alpha_i (1 - y_i (x_i^\top \beta + \beta_0)) \\
&= \frac{1}{2} \left\| \sum_{i=1}^N \alpha_i y_i x_i \right\|_2^2 + \sum_{i=1}^N \alpha_i - \left(\sum_{i=1}^N \alpha_i y_i x_i \right)^\top \left(\sum_{i=1}^N \alpha_i y_i x_i \right) \\
&= -\frac{1}{2} \left\| \sum_{i=1}^N \alpha_i y_i x_i \right\|_2^2 + \sum_{i=1}^N \alpha_i \\
&= \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i^\top x_j.
\end{aligned}$$

We find that $\alpha_i \geq 0$, $\mu_i \geq 0$ and $\alpha_i = C - \mu_i$ implies $0 \leq \alpha_i \leq C$. We therefore maximize

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i^\top x_j \quad (16)$$

$$\text{subject to } 0 \leq \alpha_i \leq C, \sum_{i=1}^N \alpha_i y_i = 0.$$

Maximizing the dual problem (16) is a simpler convex quadratic programming problem than the primal (12), and can be solved with standard techniques.

Let's check the K.K.T. conditions to gain more insight:

$$\text{Stationarity : } \begin{cases} \alpha_i = C - \mu_i \\ \sum_{i=1}^N \alpha_i y_i = 0 \\ \sum_{i=1}^N \alpha_i y_i x_i = \beta \end{cases} \quad (17)$$

$$\begin{array}{l} \text{Complementary} \\ \text{slackness} \end{array} : \begin{cases} \alpha_i (y_i (x_i^\top \beta + \beta_0) - (1 - \xi_i)) = 0 \\ \mu_i \xi_i = 0 \end{cases} \quad (18)$$

$$\begin{array}{l} \text{Primal} \\ \text{feasibility} \end{array} : \begin{cases} y_i (x_i^\top \beta + \beta_0) - (1 - \xi_i) \geq 0 \\ \xi_i \geq 0 \end{cases} \quad (19)$$

$$\begin{array}{l} \text{Dual} \\ \text{feasibility} \end{array} : \begin{cases} \alpha_i \geq 0 \\ \mu_i \geq 0 \end{cases} \quad (20)$$

From (14) we see that the solution for β has the form

$$\hat{\beta} = \sum_{i=1}^N \hat{\alpha}_i y_i x_i = \sum_{i:\alpha_i \neq 0} \alpha_i y_i x_i$$

with nonzero coefficient $\hat{\alpha}_i$ only for those observations i for which the constraints in (19) are exactly met. i.e.

- If $y_i (x_i^\top \beta + \beta_0) - (1 - \xi_i) > 0$ then $\alpha_i = 0$.
- If $\alpha_i \neq 0$, $y_i (x_i^\top \beta + \beta_0) - (1 - \xi_i) = 0$. These observations are called the **support vectors**. We have two scenarios for $\alpha_i \neq 0$, :
 - If $\xi_i = 0$, then x_i lies exactly on the margin and by (17) and (18), we get $\alpha_i \in (0, C]$.

– If $\xi_i > 0$, then by (18) $\mu_i = 0$, then by (17) $\alpha_i = C$.

Given the solutions $\hat{\beta}_0$ and $\hat{\beta}$, the decision function can be written as

$$\hat{G}(x) = \text{sgn}(x^\top \hat{\beta} + \hat{\beta}_0).$$

3 Support Vector Machines and Kernels

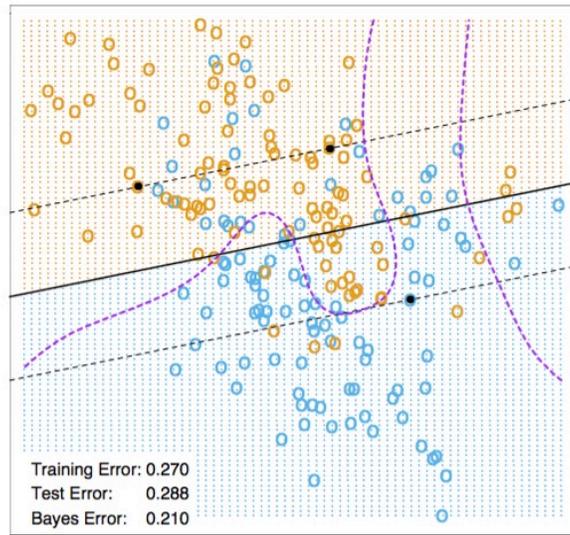
The **Lagrange dual function** has the form

$$L_D(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle,$$

and produce the classifier

$$\begin{aligned} & \text{sgn}(x^\top \beta + \beta_0) \\ &= \text{sgn}(\beta_0 + x^\top \sum_i \alpha_i y_i x_i) \\ &= \text{sgn}(\beta_0 + \sum_i \alpha_i y_i \langle x, x_i \rangle). \end{aligned}$$

so the computation and use of the fitted SVM only requires $\langle x_i, x_j \rangle$. The support vector classifier described so far finds linear boundaries in the input feature space. As with other linear methods, we can make the procedure more flexible by enlarging the feature space using basis expansions such as polynomials or splines. Now we can perform similar inner product operation for the transformed feature vectors $h(x_i)$, $h(x_i) = (h_1(x_i), h_2(x_i), \dots, h_M(x_i))$,



$C = 10000$

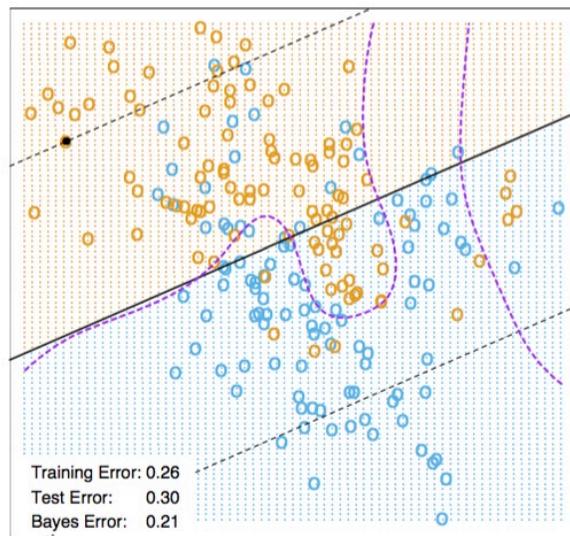


Figure 3: The linear support vector boundary for the mixture data example with two overlapping classes, for two different values of C . The broken lines indicate the margins, where $f(x) = \pm 1$. The support points ($\alpha_i > 0$) are all the points on the wrong side of their margin. The black solid 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. The broken purple curve in the background is the Bayes decision boundary.

$i = 1, \dots, n$, where each $x_i \in \mathbb{R}^p$ is transformed to $h(x_i) \in \mathbb{R}^M$.

$$L_D(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \langle h(x_i), h(x_j) \rangle,$$

Generally linear boundaries in the enlarged space achieve better training-class separation, and translate to nonlinear boundaries in the original space.

$$\begin{aligned} f(x) &= h(x)^\top \beta + \beta_0 \\ &= \sum_{i=1}^n \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0. \end{aligned}$$

In fact, we need not specify the transformation $h(x)$ at all, but require only knowledge of the kernel function

$$K(x, x') = \langle h(x), h(x') \rangle$$

that computes inner products in the transformed space. K should be a symmetric positive (semi-) definite function; Three popular choices for K in the SVM literature are

*d*th-Degree polynomial: $K(x, x') = (1 + \langle x, x' \rangle)^d$,
Radial basis: $K(x, x') = \exp(-\gamma \|x - x'\|^2)$,
Neural network: $K(x, x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2)$.

Replace $\langle x_i, x_j \rangle$ with $K(x_i, x_j)$. We require $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) > 0$.

Then SVM dual becomes

$$\begin{aligned} \max \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \\ & \sum_{i=1}^N \alpha_i y_i = 0. \end{aligned}$$

and the solution can be written as

$$\hat{f}(x) = \sum_{i=1}^N \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0.$$

Consider for example a feature space with two inputs X_1 and X_2 , and a polynomial kernel of degree 2. Then

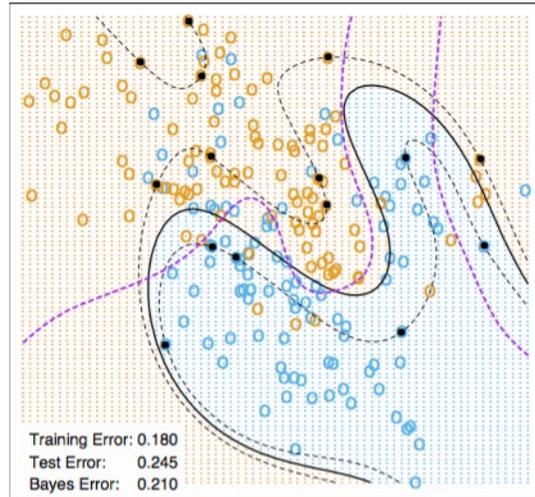
$$\begin{aligned} K(X, X') &= (1 + \langle X, X' \rangle)^2 \\ &= (1 + X_1 X'_1 + X_2 X'_2)^2 \\ &= 1 + 2X_1 X'_1 + 2X_2 X'_2 + (X_1 X'_1)^2 + (X_2 X'_2)^2 + 2X_1 X'_1 X_2 X'_2. \end{aligned}$$

Then $M = 6$, and if we chose $h_1(X) = 1$, $h_2(X) = \sqrt{2}X_1$, $h_3(X) = \sqrt{2}X_2$, $h_4(X) = X_1^2$, $h_5(X) = X_2^2$, $h_6(X) = \sqrt{2}X_1 X_2$, then $K(X, X') = \langle h(X), h(X') \rangle$.

4 The SVM as A Penalization Method

The SVM with overlapping class distributions is

SVM - Degree-4 Polynomial in Feature Space



SVM - Radial Kernel in Feature Space

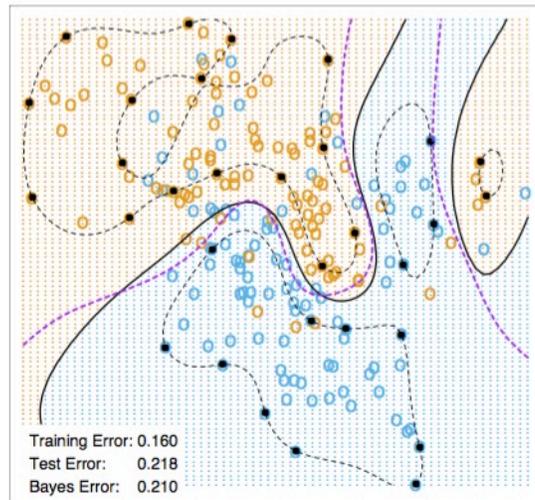


Figure 4: Two nonlinear SVMs for the mixture data. The upper plot uses a 4th degree polynomial kernel, the lower a radial basis kernel (with $\gamma = 1$). In each case C was tuned to approximately achieve the best test error performance, and $C = 1$ worked well in both cases. The radial basis kernel 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.

$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \xi_i$$

$$\text{subject to } y_i(\beta^\top x_i + \beta_0) \geq 1 - \xi_i \quad \forall i, \xi_i \geq 0$$

We can also eliminate the linear inequality constraints.

$$\begin{aligned} \forall i, \xi_i \geq 0 \text{ and } \xi_i \geq 1 - y_i(x_i^\top \beta + \beta_0) \\ \iff \max(1 - y_i(x_i^\top \beta + \beta_0), 0) \leq \xi_i. \end{aligned}$$

Given (β, β_0) , $\sum_i \xi_i$ should be minimized, therefore the optimal value ξ_i^* should have the property that $\xi_i^* = \max(1 - y_i(x_i^\top \beta^* + \beta_0^*), 0)$. Then,

$$\begin{aligned} (\beta^*, \beta_0^*) &= \arg \min_{(\beta, \beta_0)} \frac{1}{2} \|\beta\|_2^2 + C \sum_i \max(1 - y_i(x_i^\top \beta + \beta_0), 0) \\ &= \arg \min_{(\beta, \beta_0)} \sum_{i=1}^n \max(1 - y_i(x_i^\top \beta + \beta_0), 0) + \lambda \|\beta\|_2^2. \end{aligned}$$

Let $\phi_{\text{hinge}} = \max(1 - t, 0) = (1 - t)_+$ and $f(x) = x^\top \beta + \beta_0$. The problem becomes

$$\min \sum_{i=1}^N \phi_{\text{hinge}}(y_i f(x_i)) + \lambda \|\beta\|_2^2. \quad (21)$$

If the data are separable, then the limit of $\hat{\beta}_\lambda$ in (21) as $\lambda \rightarrow 0$ defines the optimal separating hyperplane. Compare it with logistic regression:

$$\min \sum_{i=1}^N \phi_{\text{logit}}(y_i f(x_i))^2 + \lambda \|\beta\|_2^2.$$

where $y_i = \pm 1$, and $\phi_{\text{logit}}(t) = \log(1 + \exp(-t))$. This error function is also plotted in Figure 5 and we see that it has a similar form to the support vector error function. The key difference is that the flat region in $\phi_{\text{hinge}}(t) = (1 - t)_+$ leads to sparse solutions.

Both the logistic error and the hinge loss can be viewed as continuous approximations to the misclassification error. Another continuous error function that has sometimes been used to solve classification problems is the squared error, which is again plotted in Figure 5. It has the property, however, of placing increasing emphasis on data points that are correctly classified but that are a long way from the decision boundary on the correct side. Such points will be strongly weighted at the expense of misclassified points, and so if the objective is to minimize the misclassification rate, then a monotonically decreasing error function would be a better choice.

5 Kernel Regression

Now we extended the model

$$\min \sum_{i=1}^N \phi(y_i, f(x_i))^2 + \lambda \|\beta\|_2^2.$$

to the kernel case. We consider the problem

$$\hat{f}(x) = \arg \min_{f \in \mathbb{H}_K} \sum_{i=1}^n \phi(y_i, f(x_i)) + \lambda \|f\|_{\mathbb{H}_K}^2, \quad (22)$$

We express $f(x)$ in a finite-dimensional subspace spanned by kernel functions on observational data, i.e.,

$$f(x) = \sum_{i=1}^n \alpha_i K(x_i, x), \quad (23)$$

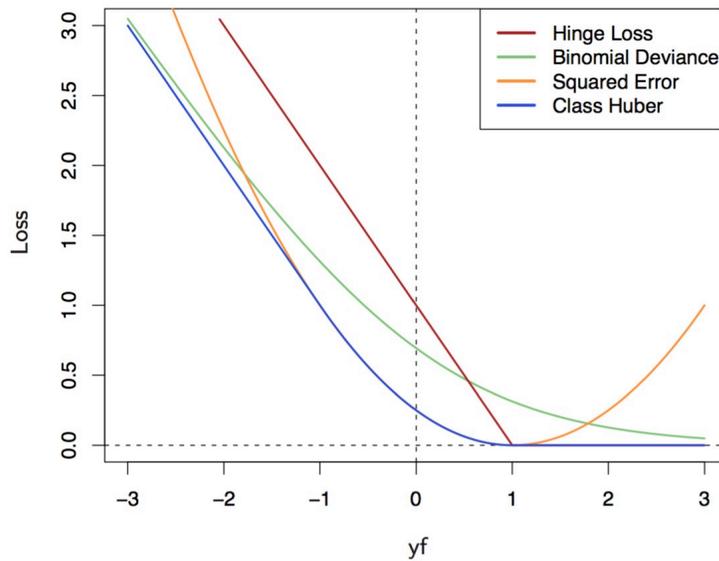


Figure 5: The support vector loss function (hinge loss), compared to the negative log-likelihood loss (binomial deviance) for logistic regression, squared-error loss, and a “Huberized” version of the squared hinge loss. All are shown as a function of yf rather than f , because of the symmetry between the $y = +1$ and $y = -1$ case. The deviance and Huber have the same asymptotes as the SVM loss, but are rounded in the interior. All are scaled to have the limiting left-tail slope of -1.

Loss Function	$L[y, f(x)]$	Minimizing Function
Binomial Deviance	$\log[1 + e^{-yf(x)}]$	$f(x) = \log \frac{\Pr(Y = +1 x)}{\Pr(Y = -1 x)}$
SVM Hinge Loss	$[1 - yf(x)]_+$	$f(x) = \text{sign}[\Pr(Y = +1 x) - \frac{1}{2}]$
Squared Error	$[y - f(x)]^2 = [1 - yf(x)]^2$	$f(x) = 2\Pr(Y = +1 x) - 1$
“Huberised” Square Hinge Loss	$-4yf(x), \quad yf(x) < -1$ $[1 - yf(x)]_+^2 \quad \text{otherwise}$	$f(x) = 2\Pr(Y = +1 x) - 1$

Figure 6: The population minimizers for the different loss functions in Figure 5. Logistic regression uses the binomial log-likelihood or deviance. Linear discriminant analysis uses squared-error loss. The SVM hinge loss estimates the mode of the posterior class probabilities, whereas the others estimate a linear transformation of these probabilities.

for some $\{\alpha_i\}_{i=1}^n \subset \mathbb{R}$. By (23) and the reproducing property of RKHS we have

$$\|f\|_{\mathbb{H}_K}^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j). \quad (24)$$

Based on (23) and (24) we can rewrite the minimization problem (22) in a finite-dimensional space

$$\{\hat{\alpha}_i\}_{i=1}^n = \arg \min_{\{\alpha_i\}_{i=1}^n} \sum_{i=1}^n \phi \left(y_i, \sum_{j=1}^n \alpha_j K(x_i, x_j) \right) + \lambda \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j). \quad (25)$$

The corresponding estimator is $\hat{f}(x) = \sum_{i=1}^n \hat{\alpha}_i K(x_i, x)$.