Support Vector Machines

STAT 37710 / CMSC 35400
Rebecca Willett and Yuxin Chen

Linearly separable training data

Consider a classification problem in which we observe training samples (x_i, y_i), $i = 1, \dots, n$, with $y_i \in \{-1, +1\}$ and the samples are

> Linearly separable — that is, there exists some (β, β_0) so that
> $$\text{sign}(\beta_0 + \beta^T x_i) = y_i \quad \text{for all } i$$

The pair (β, β_0) defines a **separating hyperplane**: $\{x : \beta_0 + \beta^T x = 0\}$

When the data is linearly separable, there is typically more than one separating hyperplane.

In this case, how should we decide which separating hyperplane to use as our predictor?

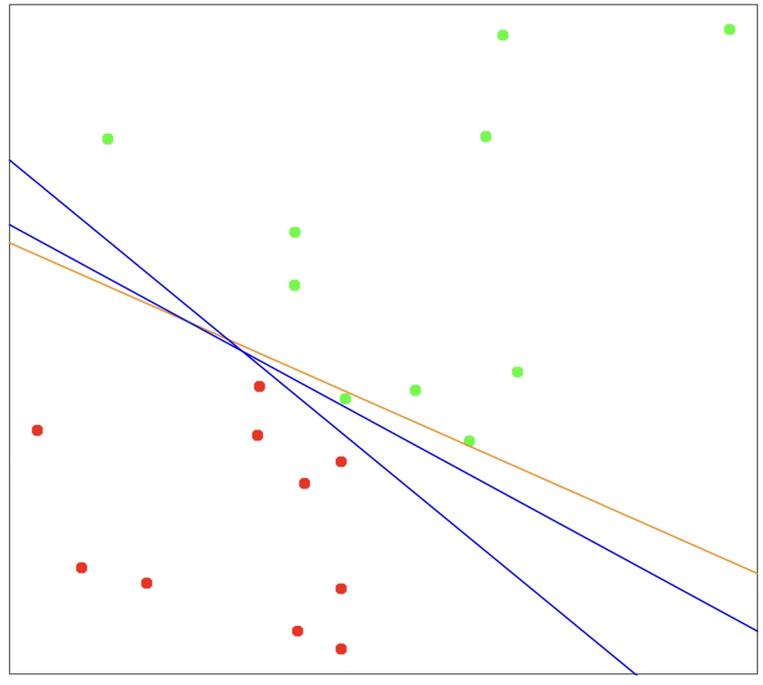
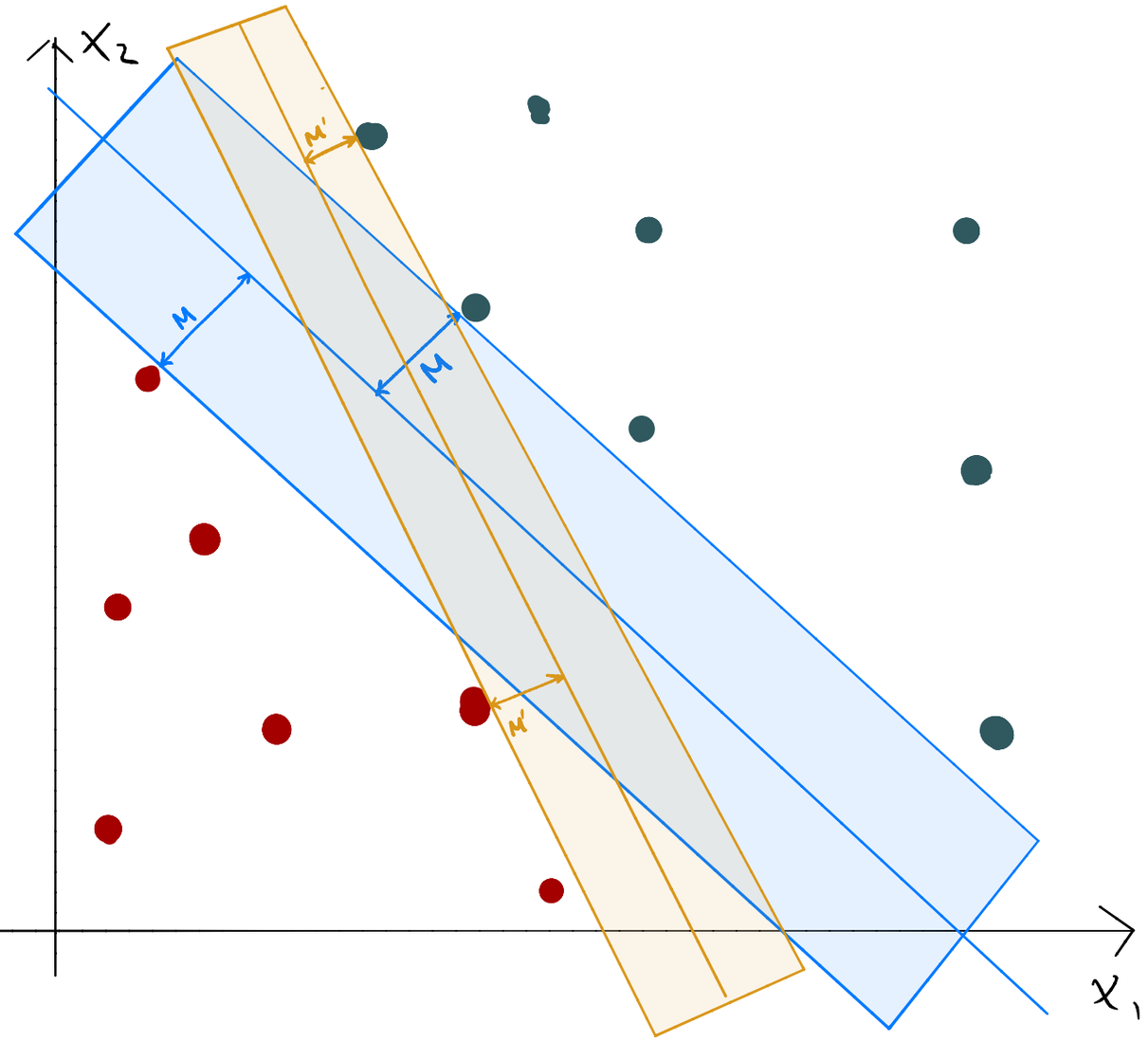


FIGURE 4.14. *A toy example with two classes separable by a hyperplane. The orange line is the least squares solution, which misclassifies one of the training points. Also shown are two blue separating hyperplanes found by the* perceptron learning algorithm *with different random starts.*

Max-margin classifiers

Proposal (Vapnik 1996): choose the separating hyperplane that maximizes the distance between the hyperplane and the closest training point

To the right, the blue hyperplane is at least distance M from each sample, whereas the gold hyperplane is at least $M' < M$ from each sample. We prefer the blue $\Rightarrow$ better performance on test data.



Mathematically, we want to choose β, β_0 via

$$\max_{\beta, \beta_0 : \|\beta\|_1} M \quad \text{subject to } y_i(\beta_0 + \beta^T x_i) \geq M \quad \text{for } i = 1, \ldots, n$$

which is equivalent to

$$\max_{\beta, \beta_0} M \quad \text{subject to } y_i(\beta_0 + \beta^T x_i) \geq M\|\beta\|$$

Now set $\|\beta\| = 1/M$ (since any rescaling of (β, β_0) makes the inequalities equivalent), yielding

$$\max_{\beta, \beta_0} \tfrac{1}{2}\|\beta\|^2 \quad \text{subject to } y_i(\beta_0 + \beta^T x_i) \geq 1, \quad i = 1, \ldots, n$$

Optimization problem and dual

How can we solve this optimization problem? First, write it in the Lagrange form:

$$L_p := \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\beta_0 + \beta^T x_i) - 1 \right] \quad \text{for Lagrange multipliers } \alpha_i \geq 0$$

The gradient of this with respect to β is

$$\nabla_\beta L_p = \beta - \sum_{i=1}^{n} \alpha_i y_i x_i, \quad \frac{dL_p}{d\beta_0} = -\sum_{i=1}^{n} \alpha_i y_i$$

Set derivatives to zero to get

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i, \quad 0 = \sum_{i=1}^{n} \alpha_i y_i$$

Plug these back into L_p to get

$$\frac{1}{2}\sum_{j=1}^{p}\left(\sum_{i=1}^{n} \alpha_i y_i x_{ij}\right)^2 - \sum_{i=1}^{n} \alpha_i \left[y_i \left(\beta_0 + \left[\sum_{k=1}^{n} \alpha_k y_k x_k\right]^T x_i\right) - 1 \right]$$

$$= \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k x_{ij} x_{ik} - \sum_{i=1}^{n} \alpha_i y_i \beta_0 - \sum_{i=1}^{n} \alpha_i y_i \left(\sum_{k=1}^{n} \alpha_k y_k x_k\right)^T x_i + \sum_{i=1}^{n} \alpha_i$$

($\sum_{i=1}^{n} \alpha_i y_i \beta_0 = 0$)

$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k x_i^T x_k + \sum_{i=1}^{n} \alpha_i$$

Now we can get values for the α's by solving the "dual" problem

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n} \alpha_i \alpha_k y_i y_k x_i^T x_k \quad \text{subject to } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0$$

Optimization theory shows us that the optimal solution will satisfy

$$\alpha_i \left[y_i(\beta_0 + \beta^T x_i) - 1 \right] = 0$$

for all i.

- if $\alpha_i > 0$, then
  $y_i(\beta_0 + \beta^T x_i) = 1$
  $\Rightarrow x_i$ is on boundary of the margin

- if $y_i(\beta_0 + \beta^T x_i) > 1$, then
  $\alpha_i = 0$ and x_i is not on the boundary of the margin.

x_i's for which $\alpha_i > 0$ are called support vectors

Support vectors

Key implication: optimal separating hyperplane only depends on the support vectors, points closest to the decision boundary. This is very different than the least squares solution, which depends on all the training data

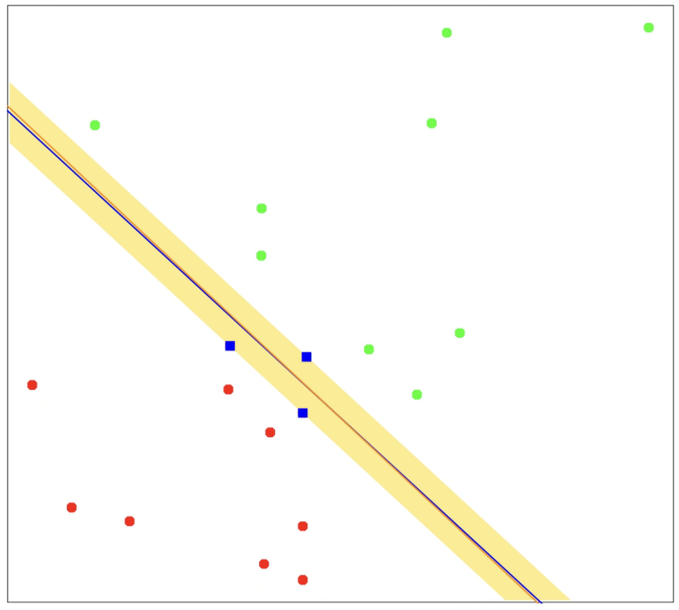


FIGURE 4.16. *The same data as in Figure 4.14. The shaded region delineates the maximum margin separating the two classes. There are three support points indicated, which lie on the boundary of the margin, and the optimal separating hyperplane (blue line) bisects the slab. Included in the figure is the boundary found using logistic regression (red line), which is very close to the optimal separating hyperplane (see Section 12.3.3).*

Inseparable data

Now suppose there is NO linearly separating hyperplane.

Approach: still maximize M, but let some points be on the wrong side.

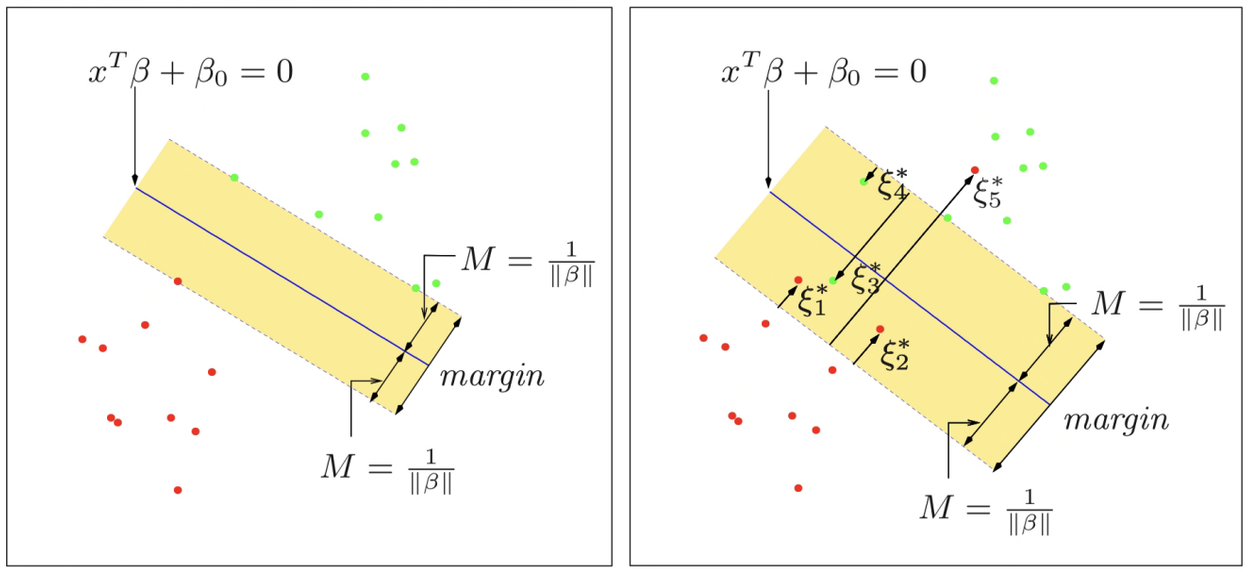


FIGURE 12.1. *Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width $2M = 2/\|\beta\|$. The right panel shows the nonseparable (overlap) case. The points labeled ξ_j^* are on the wrong 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 $\sum \xi_i \leq$ constant. Hence $\sum \xi_j^*$ is the total distance of points on the wrong side of their margin.*

Let $\xi_i = \frac{1}{M} \cdot$ how far x_i is from being on the wrong side of the margin. each $\xi_i \geq 0$

(where $\xi_i = 0$ implies x_i is on the correct side of the margin).

Now instead of

maximizing M subject to $y_i(\beta_0 + \beta^T x_i) \geq M$

we

maximize M subject to $y_i(\beta_0 + \beta^T x_i) \geq M(1-\xi_i)$

and $\xi_i \geq 0$ and $\sum_{i=1}^{n} \xi_i \leq 1/\lambda$

Putting the pieces together, we arrive at the specification of a support vector classifier:

$$\min_{\beta, \beta_0, \xi_i} \frac{1}{2}\|\beta\|^2 \quad \text{subject to} \quad y_i(\beta_0 + \beta^T x_i) \geq 1 - \xi_i \quad \text{for all } i$$
$$\xi_i \geq 0, \quad \sum_i \xi_i \leq 1/\lambda$$

Dual formulation and optimal β

From the previous page, we had

$$\min_{\beta, \beta_0, \xi_i} \frac{1}{2}\|\beta\|^2 \quad \text{subject to} \quad y_i(\beta_0 + \beta^T x_i) \geq 1 - \xi_i \text{ \& } \xi_i \geq 0 \quad \text{for all } i \quad \sum_i \xi_i \leq 1/\lambda$$

This is equivalent to

$$\min_{\beta, \beta_0, \xi_i} \sum_{i=1}^n \xi_i + \frac{\lambda}{2}\|\beta\|^2 \quad \text{subject to } \xi_i \geq 0 \text{ \& } y_i(\beta_0 + \beta^T x_i) \geq 1 - \xi_i \text{ for all } i$$

Again using Lagrange multipliers, we arrive at the dual problem

$$\max_{\alpha_i} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{k=1}^n \alpha_i \alpha_k y_i y_k \underline{x_i^T x_k} \quad \text{subject to } 0 \leq \alpha_i \leq 1/\lambda \text{ and } \sum_{i=1}^n \alpha_i y_i = 0 \qquad (*)$$

In practice, we solve this problem for the α_i's, and the optimal $\beta = \sum_{i=1}^n \alpha_i y_i x_i$.

Kernelized SVMs

Note that ⊛ only uses the inner products of feature vectors — this means we can apply the kernel trick!

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k \underbrace{K(x_i, x_k)}_{\phi(x_i)^T \phi(x_k)}$$

$$\text{subject to } 0 \leq \alpha_i \leq 1/\lambda \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

The predictor $\beta_0 + \beta^T \phi(x)$ can then be expressed as

$$\beta_0 + \beta^T \phi(x) = \beta_0 + \left(\sum_{i=1}^{n} \alpha_i y_i \phi(x_i) \right)^T \phi(x)$$

$$= \beta_0 + \sum_{i=1}^{n} \alpha_i y_i \underbrace{K(x_i, x)}_{\phi(x_i)^T \phi(x)}$$

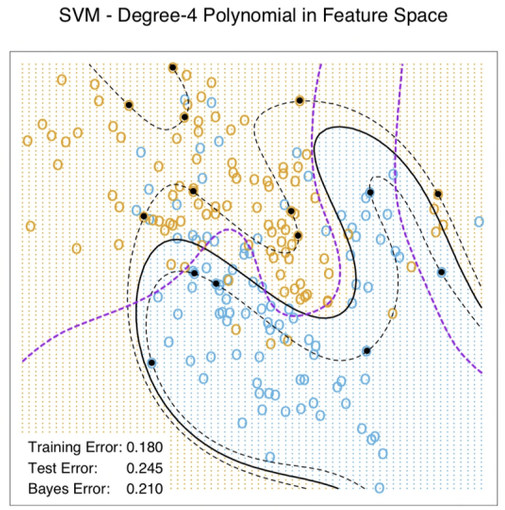


FIGURE 12.3. *Two nonlinear SVMs for the mixture data. The upper plot uses a 4th degree polynomial kernel, the lower a radial basis kernel (with $\gamma = 1$). In each case C was tuned to approximately achieve the best test error performance, and $C = 1$ worked well in both cases. The radial basis kernel 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.*

Hinge loss perspective

Let $\hat{y} = f(x) = \beta_0 + \beta^T \phi(x)$, and consider

Choosing β via

$$\min_{\beta_0, \beta} \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \|\beta\|^2$$

"hinge loss" — penalty

The solution to this is the same as the solution to

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

subject to $\xi_i \geq 0$ & $y_i(\beta_0 + \beta^T x_i) \geq 1 - \xi_i$ for all i

This classifier is referred to as a

SUPPORT VECTOR MACHINE

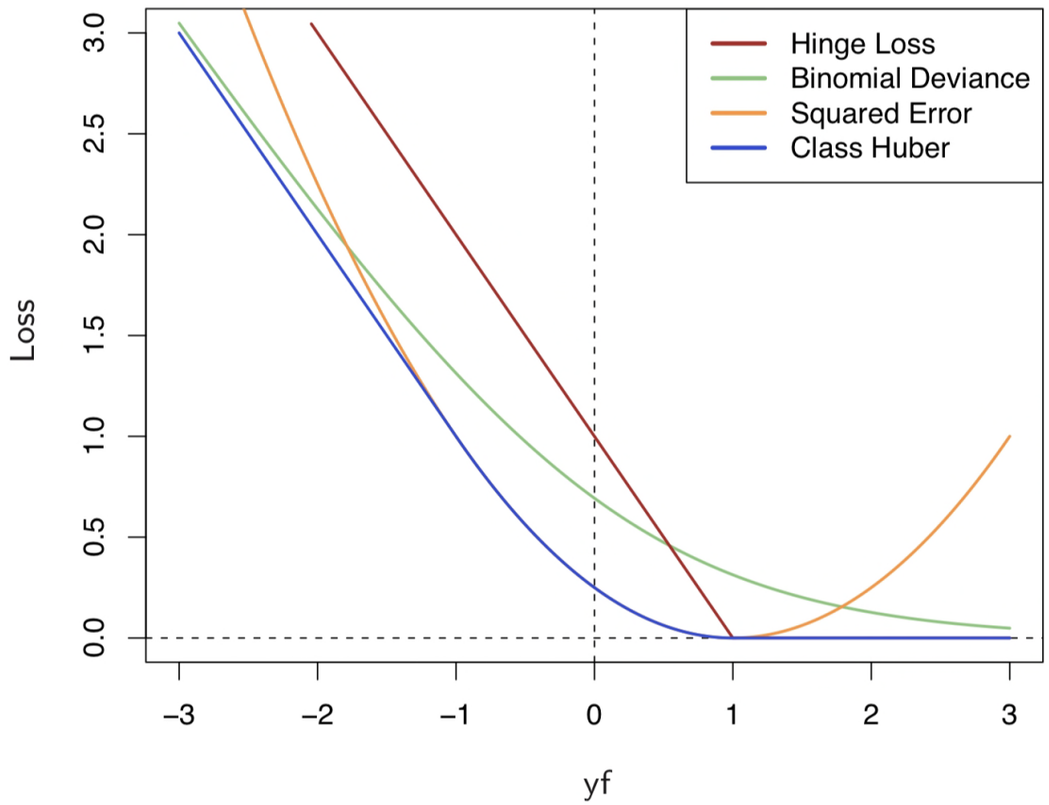


FIGURE 12.4. *The support vector loss function (hinge loss), compared to the negative log-likelihood loss (binomial deviance) for logistic regression, squared-error loss, and a "Huberized" version of the squared hinge loss. All are shown as a function of yf rather than f, because of the symmetry between 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.*