Outlines

- Soft and Hard Classifiers
- Separating Hyperplane
  - Binary SVM for Separable Case
  - Linear SVM for Non-separable Problems
  - Nonlinear SVM
Soft and Hard Classifiers

- **Soft** classifiers estimate class (posterior) probabilities first and then make a decision:
  1. first estimate the conditional probability $P(Y = +1|X = x)$;
  2. then construct the decision rule as $\hat{P}(Y = +1|x) > 0.5$.

Examples: LDA, Logistics regression, knn methods.

- **Hard** classifiers do not estimate the class probabilities. Instead, they target on the Bayes rule directly:
  - estimate $I[\hat{P}(Y = +1|x) > 0.5]$ directly.

Examples: SVM
Construct linear decision boundary that explicitly tries to separate the data into different classes as well as possible.

Good separation is defined in a certain form mathematically.

Even when the training data can be perfectly separated by hyperplanes, LDA and other linear methods developed under a statistical framework may not achieve perfect separation.

For convenience, we label two classes as $+1$ and $-1$ from now on.
Figure 4.13: 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.
For convenience, we define

\[ f(x) = \beta_0 + \beta^T x. \]

Note that

- If a point is correctly classified, then its signed distance has the same sign as its label. It implies that
  \[ y_i f(x_i) > 0. \]

- If a point is misclassified, then its signed distance has the opposite sign as its label. It implies that
  \[ y_i f(x_i) < 0. \]

Call the product \( y_i f(x_i) \) the functional margin of the classifier \( f \).
Main motivation:

Separate two classes and maximizes the distance to the closest points from either class (Vapnik 1996)

- Provides a unique solution
- Leads to better classification performance on test (future) data
Figure 4.15: The same data as in Figure 4.13. 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).
Let \( f(x) = \beta_0 + x^T \beta \). Consider the problem:

\[
\max_{\beta_0, \beta, C} C \\
\text{subject to } \frac{y_i(x_i^T \beta + \beta_0)}{\|\beta\|} \geq C, \quad i = 1, \ldots, n.
\]

All the points are correctly satisfied, and they have at least a signed distance \( C \) from the separating hyperplane.

- The maximal \( C \) must be positive (for separable data).
- The quantity \( yf(x) \) is called the functional “margin”.
- A positive “margin” implies a correct classification on \( x \).

The goal is seek the largest such \( C \) and associated parameters.
The above problem is equivalent to This lead to

$$\min_{\beta_0, \beta} \|\beta\|$$

subject to

$$y_i(x_i^T \beta + \beta_0) \geq 1, \quad i = 1, \ldots, n.$$

For computational convenience, we further replace $\|\beta\|$ by $\frac{1}{2} \|\beta\|^2$ (a monotonic transformation), which leads to

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

subject to

$$y_i(x_i^T \beta + \beta_0) \geq 1, \quad \text{for} \quad i = 1, \ldots, n$$

This is the linear SVM for perfectly separable cases.
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 $\xi_j^*$ 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.
Solve SVM by Quadratic Programming

Need to solve a convex optimization problem:

- quadratic objective function
- linear inequality constraints

Lagrange (primal) function: introduce Lagrange multipliers $\alpha_i \geq 0$

$$L_P(\beta, \beta_0, \alpha) = \frac{1}{2} \| \beta \|^2 - \sum_{i=1}^{n} \alpha_i [y_i(x_i^T \beta + \beta_0) - 1]$$

$L$ has to be minimized with respect to the *primal variables* $(\beta, \beta_0)$ and maximized with respect to the *dual variables* $\alpha_i$. (In other words, we need to find the saddle points for $L_P$).
Solve the Wolfe Dual Problem

At the saddle points, we have

\[ \frac{\partial}{\partial \beta_0} L(\beta, \beta_0, \alpha) = 0, \quad \frac{\partial}{\partial \beta} L(\beta, \beta_0, \alpha) = 0 \]

i.e.

\[ \begin{align*}
0 &= \sum_i \alpha_i y_i, \\
\beta &= \sum_{i=1}^n \alpha_i y_i x_i.
\end{align*} \]

By substituting both into \( L \) to get the dual problem:

\[ L_D(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \]

subject to

\[ \alpha_i \geq 0, \quad i = 1, \ldots, n; \quad \sum_{i=1}^n \alpha_i y_i = 0. \]
First, solve the dual problem for $\alpha$.

$$
\min_{\alpha} \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{n} \alpha_i,
$$

subject to $\alpha_i \geq 0$, $i = 1, \ldots, n$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. Denote the solution as $\hat{\alpha}_i$, $i = 1, \ldots, n$.

The SVM slope is obtained by $\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i$.

The SVM intercept $\hat{\beta}_0$ is solved from KKT condition

$$
\hat{\alpha}_i [y_i (x_i^T \hat{\beta} + \hat{\beta}_0) - 1] = 0,
$$

with any of the points with $\hat{\alpha} > 0$ (support vectors)

The SVM boundary is given by $\hat{f}(x) = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i^T x$. 
The optimal solution should satisfy the following:

\[ 0 = \sum_{i} \alpha_i y_i \]  \hspace{1cm} (1)

\[ \beta = \sum_{i=1}^{n} \alpha_i y_i x_i \]  \hspace{1cm} (2)

\[ \alpha_i \geq 0, \quad \text{for } i = 1, \ldots, n \]  \hspace{1cm} (3)

\[ 0 = \sum_{i=1}^{n} \alpha_i [y_i (x_i^T \beta + \beta_0) - 1], \]  \hspace{1cm} (4)

\[ 1 \leq y_i (x_i^T \beta + \beta_0), \quad \text{for } i = 1, \ldots n. \]  \hspace{1cm} (5)
Optimality condition (4) is called the “complementary slackness”

\[ 0 = \sum_{i=1}^{n} \alpha_i [y_i(x_i^T \beta + \beta_0) - 1]. \]

Since \( \alpha_i \geq 0 \) and \( y_if(x_i) \geq 1 \) for all \( i \), this implies that the SVM solution must satisfy

\[ \alpha_i [y_i(x_i^T \beta + \beta_0) - 1] = 0, \quad \forall i = 1, \ldots, n. \]

In other words,

- If \( \alpha_i > 0 \), then \( y_i(x_i^T \beta + \beta_0) = 1 \), so \( x_i \) must lie on the margin.
- If \( y_i(x_i^T \beta + \beta_0) > 1 \) (not on the margin), then \( \alpha_i = 0 \).

No training points fall in the margin.
Support Vectors (SVs): all the points with $\alpha_i > 0$. Define the index set

$$SV = \{i | \hat{\alpha}_i > 0, \ i = 1..., n\}$$

- The solution and decision boundary can be expressed only in terms of SVs
  
  $$\hat{\beta} = \sum_{i \in SV} \hat{\alpha}_i y_i x_i$$
  
  $$\hat{f}(x) = \hat{\beta}_0 + \sum_{i \in SV} \hat{\alpha}_i y_i x_i^T x.$$  

- The identification of the SV points requires the use of all the data points.
The linear SVM for perfectly separable cases:

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

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

For non-separable data,

- Such a pair \((\beta_0, \beta)\) does not exist!
- Why?
- How to generalize the linear SVM here?
Relaxed constraints by introducing positive variables $\xi_i \geq 0$ for each $i$.

$$x^T \beta + \beta_0 \geq 1 - \xi_i \quad \text{for } y_i = 1$$
$$x^T \beta + \beta_0 \leq -1 + \xi_i \quad \text{for } y_i = -1$$

Basic ideas:

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

- Similar for the violations in $-1$ class.

The above two inequalities can be summarized as

$$y_i(x^T \beta + \beta_0) \geq 1 - \xi_i, \quad \forall i = 1, \cdots, n.$$
Linear SVM for non-separable problems:

\[
\min_{\beta, \beta_0, \xi} \quad \frac{1}{2}||\beta||^2 + \gamma(\sum_{i=1}^{n} \xi_i)
\]
subject to \( y_i(x^T \beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \ldots, n. \)
\( \xi_i \geq 0, \quad i = 1, \ldots, n, \)

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

- If an error occurs at the training point \((x_i, y_i)\), 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 $\xi_j^*$ 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 $\sum \xi_i \leq \text{constant}$. Hence $\sum \xi_j^*$ is the total distance of points on the wrong side of their margin.
We say $\gamma$ the *regularization*, or tuning, or cost parameter.

- $\gamma$ 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

$$
\min_{\beta, \beta_0, \xi} \quad \frac{1}{2}||\beta||^2 + C \left( \sum_{i=1}^{n} \xi_i \right)
$$

subject to

$$
y_i(x^T\beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \ldots, n.
$$

$$
\xi_i \geq 0, \quad i = 1, \ldots, n.
$$

The SVM optimization is a QP.
Introducing the Lagrange multipliers $\alpha_i \geq 0, \mu_i \geq 0$ for each constraint, we obtain the Lagrange function

$$L(\beta, \beta_0, \alpha, \xi, \mu) \equiv \frac{1}{2}||\beta||^2 + \gamma \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \mu_i \xi_i$$

$$- \sum_{i=1}^{n} \alpha_i [y_i(\beta^T x_i + \beta_0) - 1 + \xi_i]$$

Optimization problems:

- minimize $L$ with respect to primal variables $\beta, \beta_0, \xi$
- maximize $L$ with respect to dual variables $\alpha, \mu$. 
Using the KKT theory, the optimal solution should satisfy the following equations

\[
\frac{\partial}{\partial \beta_0} L = 0, \quad \frac{\partial}{\partial \beta} L = 0, \quad \frac{\partial}{\partial \xi_i} L = 0.
\]

We get

\[
0 = \sum_{i=1}^{n} \alpha_i y_i
\]

\[
\beta = \sum_{i=1}^{n} \alpha_i y_i x_i
\]

\[
\alpha_i = \gamma - \mu_i, \quad i = 1, \ldots, n
\]
KKT Conditions for Soft-margin SVM

- Stationary conditions:
  \[0 = \sum_{i=1}^{n} \alpha_i y_i, \quad \beta = \sum_{i=1}^{n} \alpha_i y_i x_i, \quad \alpha_i = \gamma - \mu_i, \quad i = 1, ..., n.\]

- Primal feasible:
  \[y_i(x^T \beta + \beta_0) - (1 - \xi_i) \geq 0, \quad \xi_i \geq 0, \quad i = 1, ..., n.\]

- Dual feasible:
  \[\alpha_i \geq 0, \quad \mu_i \geq 0, \quad i = 1, ..., n.\]

- Complementary slackness:
  \[0 = \alpha_i [y_i(\beta^T x_i + \beta_0) - 1 + \xi_i], \quad 0 = \mu_i \xi_i, \quad i = 1, ..., n.\]

They together **uniquely** characterize the solution to primal/dual problem.
If $y_i(\beta^T x_i + \beta_0) > 1$ (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 $y_i(\beta^T x_i + \beta_0) = 1$ (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 $y_i(\beta^T x_i + \beta_0) < 1$ (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)
The set of support vectors: $SV = \{i | \hat{\alpha}_i > 0, \ i = 1..., \ n\}$

$$\hat{\beta} = \sum_{i \in SV} \hat{\alpha}_i y_i x_i$$

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 $\alpha_i = \gamma$. 
By substituting both into $L_P$, we get

$$G(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

subject to

$$0 \leq \alpha_i \leq \gamma, \quad i = 1, \ldots, n;$$

$$\sum_{i} \alpha_i y_i = 0.$$

Maximizing the dual is a simpler convex QP than the primal.
Obtain SVM Primal Solutions

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

$$\beta^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i.$$ 

To solve $\beta_0$
- Use any point satisfying $\alpha_i > 0, \xi_i = 0$.
- Typically use an average of all the solutions for numerical stability.
Tuning Parameter $\gamma$

- Large $\gamma$ puts more weight on misclassification rate than margin width
  - discourage any positive $\xi_i$
  - may lead to an overfit wiggly boundary in the original space
- Small $\gamma$ 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
$\gamma = 0.01$
Figure 12.2: The linear support vector boundary for the mixture data example with two overlapping classes, for two different values of $\gamma$. 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

- Data cannot 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
Nonlinearly map data into some other inner product space

\[ \Phi : \mathbb{R}^d \rightarrow \mathcal{F} \]

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

\[
h(x_i) = (h_1(x_i), h_2(x_i), ..., h_M(x_i))
\]

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

The dimension of the enlarged space can have very large, infinite dimensions.
The computation can be prohibitive.
Support Vector Classifiers

input space

feature space

\[ \Phi \]

Hao Helen Zhang
Lecture 13: Support Vector Machines

B. Schölkopf, Canberra, February 2002
The Lagrange dual function

\[ G(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j < h(x_i), h(x_j) > \]

subject to

0 ≤ \alpha_i ≤ \gamma, \quad i = 1, \ldots, n; \quad \sum_i \alpha_i y_i = 0.

The solution function

\[ f(x) = h(x)^T \beta + \beta_0 = \sum_{i=1}^{n} \alpha_i y_i < h(x), h(x_i) > + \beta_0. \]

Given \alpha_i, the intercept \beta_0 is determined by solving \( f(x_i) = 0 \) for any \( x_i \) for which 0 < \alpha_i < \gamma.
The prediction function only relies on $< h(x), h(x_i) >$. We define this inner product at kernel function

$$K(x, x') = < h(x), h(x') >$$

- $K$ symmetric positive (semi-) definite function
- Various kernels are used in literature:
  - $d$th Degree polynomial: $K(x, x') = (1 + < x, x' >)^d$
  - Radial basis: $K(x, x') = \exp(-\|x - x'\|^2/\sigma)$
  - Neural network: $K(x, x') = \tanh(\kappa_1 < x, x' > + \kappa_2)$
For $d = 2$, the kernel function value $(x_1, x_2)$ and $(x'_1, x'_2)$

$$K(x, x') = (1 + \langle x, x' \rangle)^2$$
$$= (1 + x_1x'_1 + x_2x'_2)^2$$
$$= 1 + 2x_1x'_1 + 2x_2x'_2 + (x_1x'_1)^2 + (x_2x'_2)^2 + 2x_1x'_1x_2x'_2$$

Then $M = 6$, with

$$h_1(x) = 1 \quad h_2(x) = \sqrt{2}x_1 \quad h_3(x) = \sqrt{2}x_2$$
$$h_4(x) = x_1^2 \quad h_5(x) = x_2^2 \quad h_6(x) = \sqrt{2}x_1x_2$$

Check: $K(x, x') = \langle h(x), h(x') \rangle$. 
Example: All Degree 2 Monomials

\[ \Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3 \]
\[ (x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2} x_1 x_2, x_2^2) \]
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 $\gamma$ 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.
SVM - Radial Kernel in Feature Space

Training Error: 0.160
Test Error:       0.218
Bayes Error:    0.210
R Functions: SVM

- R package *kernlab*; function *ksvm*.
- R package *e1071*; function *svm*.
- R package *svmpath*: compute the entire regularized solution path.
If the classes are really Gaussian, then
- the LDA is optimal.
- the separating hyperplane pays a price for focusing on the (noisier) data at the boundaries

Optimal separating hyperplane has less assumptions, thus more robust to model misspecification.
- In the separable case, logistic regression solution is similar to the separating hyperplane solution.
- For perfectly separable case, the log-likelihood can be driven to zero.