Support Vector Machine
Soft-Margin SVM: Primal & Dual

M. R. Hasan
Readings

• Alpaydin: 13.3
• Murphy: 14.5.2.1, 14.5.2.4, 14.5.3
• Geron: chapter 5, appendix C
What We Will Cover

• SVM: Soft margin classification
• Slack variable
• Soft margin primal & dual problems
• Upper bound for the error
• Implementation Issues
Linearly Separable Data

No outlier:

Linearly Non-Separable Data

Outlier:
Linearly Separable Data

Vapnik & Chervonenkis, 1963

No outlier:
Hard Margin Classifier

Outlier:
Soft Margin Classifier

Corinna Cortes & Vapnik, 1993

\[
\begin{align*}
\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \\
\text{subject to } y_i(w^T \tilde{x}_i + b) \geq 1 - \xi_i \quad \forall i \\
\xi_i \geq 0
\end{align*}
\]

\[
\begin{align*}
\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{N} h(y_i(w^T \tilde{x}_i + b)) \\
h(z) = \max \{0, 1 - z\}
\end{align*}
\]
Linearly Non-Separable Data

Feature Augmentation: Map data onto a higher-dimensional space

Increased dimension of the feature space increases the complexity!

As a solution use Kernel SVM

Solve the margin maximization problem on a high-dimensional feature space without actually adding the features (kernel trick).

Bernhard Boser, Isabelle Guyon and Vladimir Vapnik, 1992
Linearly Non-Separable Data

Kernel SVM

Solve the margin maximization problem on a high-dimensional feature space without actually adding the features (kernel trick).

Polynomial Kernel: new features are the polynomial and interaction terms of the existing features.

Gaussian Radial Basis Function (RBF) Kernel: new features are created based on similarity with the existing features.
Linearly Separable Data

Hard Margin Classifier

No outlier: Vapnik & Chervonenkis, 1963

Outlier:

Soft Margin Classifier

Corinna Cortes & Vapnik, 1993

Linearly Non-Separable Data

Kernel SVM

Bernhard Boser, Isabelle Guyon and Vladimir Vapnik, 1992

Corinna Cortes & Vapnik, 1993
SVM: Linear Classifier - Soft Margin

Linearly Separable Data

Outlier:

Corinna Cortes & Vapnik, 1993
Soft Margin SVM: Primal & Dual Problem
Summary in 8 slides
Soft Margin SVM: Primal Problem

• If we cannot find a hard margin classifier, then **allow some errors**.

• Cut **some “slack”** to the violators.

We allow every data point to **violate** the constraint by some slack $\xi_i$. 
The slack variables $\xi_i \geq 0$ store the deviation from the margin.

There are two types of deviation:

- An instance may lie on the wrong side of the hyperplane and be misclassified (- point).
- Or, it may be on the right side but may lie on (or inside) the margin, namely, not sufficiently away from the hyperplane (+ point).
Soft Margin SVM: Primal Problem

• Therefore, to accommodate these *errors*, our *updated decision rule* is:

\[ y_i(\overline{\mathbf{w}}^T \overline{x}_i + b) \geq 1 - \xi_i \quad \forall i \]

Example: the “+” point is not off the street: 
\((\mathbf{w}^T \mathbf{x} + b)\) is not greater than 1.

However, for some \(\xi_i\), we will allow the “+” point to be *classified correctly*. 
• Consider some cases for various $\xi_i$.

(a) $\xi_i = 0$. The point is on the **correct side** and far away from the margin.

(b) $\xi_i = 0$. The point is on the **correct side** and on the margin.

(c) $\xi_i > 0$. The point is on the correct side, **but is on the margin and not sufficiently away**.

(d) $\xi_i > 1$. The point is misclassified (above the decision boundary).

\[ y_i (\overrightarrow{w}^T \overrightarrow{x_i} + b) \geq 1 - \xi_i \quad \forall i \]

The number of **non-separable points** is $\#\{\xi_i > 0\}$, e.g., “c”, “d”
Soft Margin SVM: Primal Problem

• In the soft margin case our **optimization** goal is **two fold**:
  - To maximize the margin (i.e., \( \min_{\overline{w},b} \frac{1}{2} \overline{w}^T \overline{w} \))
  - **Minimize non-separable points** (\( \#\{\xi_i > 0\} \)).
• Based on this new consideration we **update the minimization problem** as follows.

\[
\min_{\overline{w},b} \frac{1}{2} \|\overline{w}\|^2 \equiv \frac{1}{2} \overline{w}^T \overline{w}
\]

**subject to** \( y_i(\overline{w}^T \overline{x}_i + b) \geq 1 \ \forall i \)
Soft Margin SVM: Primal Problem

- The new **formulation** of the **minimization problem** for the **soft margin** SVM classification includes the **slack variable** $\xi$.

\[
\begin{align*}
\min_{\vec{w}, b, \xi_i} & \quad \frac{1}{2} \vec{w}^T \vec{w} + C \sum_{i=1}^{N} \xi_i \\
\text{subject to} & \quad y_i (\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i \quad \forall i \\
& \quad \xi_i \geq 0
\end{align*}
\]

- Maximize the margin by **cutting some slack** to the violators.

- We can control the amount of slack we want to cut for the violators via "C".

- **C** is the **tradeoff parameter**.

- Smaller **C**: cut more slack.

- Larger **C**: No slack, be **strict to the violators**.
Soft Margin SVM: Primal Problem

- Here, $C$ is the **penalty factor**.
- It is **similar to the regularization ($l_2$ norm)** of the weight vector in linear and logistic regression.
- Note that we are penalizing not only the misclassified points but also the ones in the margin for better generalization.

$$\min_{\vec{w}, b, \xi_i} \frac{1}{2} \vec{w}^T \vec{w} + C \sum_{i=1}^{N} \xi_i$$

subject to $y_i(\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i \quad \forall i$

$\xi_i \geq 0$
Soft Margin SVM: Dual Problem

**Primal**

$$\min_{\vec{w}, b, \xi_i} \frac{1}{2} \vec{w}^T \vec{w} + C \sum_{i=1}^{N} \xi_i$$

subject to

$$y_i(\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0$$

**Dual**

$$\mathcal{L}(\vec{w}^*, b^*, \alpha, \xi_i) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j$$

subject to

$$\sum_{i=1}^{N} \alpha_i^* y_i = 0$$

$$\vec{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \vec{x}_i$$

$$0 \leq \alpha_i \leq C \quad \forall i$$
Soft Margin SVM: Primal & Dual Problem

Long Story

Linearly Separable Data

Outlier:

Corinna Cortes & Vapnik, 1993
Linear SVM Classification

- There would be **two possible scenarios** in linear SVM classification.
- In our decision rule, we can **strictly impose** that all instances be off the street and on the right side.
- This is called **hard margin classification**.

The flexible version of it is called **soft margin classification**.
Linear SVM Classification

- There are **two main issues** with hard margin classification.
  - First, it only works if the data is **linearly separable**.
  - Sometimes data could be **linearly non-separable** even after applying the kernel trick, in which case hard margin classification fails to find a decision boundary.
Linear SVM Classification

- There are **two main issues** with hard margin classification.
- Second, it is quite **sensitive to outliers**.
- Following figure shows the *Iris* dataset with just **one additional outlier**.

It is **impossible** to find a hard margin for this case.
Linear SVM Classification

- **Another example**: the decision boundary (right) ends up very different from the one *(left) without the outlier.*

- This decision boundary (right) will probably not **generalize as well**.
Linear SVM Classification

- To avoid these issues it is preferable to use a **more flexible model**.
- The objective is to find a **good balance** between keeping the street as large as possible and limiting the **margin violations**.
- I.e., instances end up in the middle of the street or even on the wrong side.
- This is called **soft margin classification**.

![Graph](image.png)
Soft Margin SVM Classification

- If the two classes are *not linearly separable* (even after using the kernel trick) such that there is no hyperplane to separate them, then we need to **allow some errors**.
- What we do is look for a solution (margins) the one that **incurs the least error**.
- In other words, we **allow some “slack”**.

We allow every data point to **violate** the constraint by some slack $\xi_i$. 
Soft Margin SVM Classification

- The **slack variables** $\xi_i \geq 0$ store the deviation from the margin.
- There are **two types of deviation**:

  Or, it may be on the right side but **may lie on (or inside) the margin**, namely, not sufficiently away from the hyperplane (+ point).

An instance may lie on the **wrong side of the hyperplane** and be misclassified (- point)
Soft Margin SVM Classification

- Therefore, to accommodate these errors, our updated decision rule is:

\[ y_i(\overrightarrow{w}^T \overrightarrow{x}_i + b) \geq 1 - \xi_i \quad \forall i \]

Example: the “+” point is not off the street: 
\((w^T x + b)\) is not greater than 1.

However, for some \(\xi_i\), we will allow the “+” point to be classified correctly.
Soft Margin SVM Classification

• Consider some cases for \( \xi_i \).

(a) \( \xi_i = 0 \).
So, \( y_i (w^T x_i + b) > 1 \)
The point is on the correct side and far away from the margin

(b) \( \xi_i = 0 \).
The point is on the correct side and on the margin

\[
y_i (\overline{w}^T \overline{x}_i + b) \geq 1 - \xi_i \quad \forall i
\]
Soft Margin SVM Classification

• Consider some cases for various \( \xi_i \).

\[
y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i
\]

(c) For point c: \( 0 < \xi_i < 1 \)

Thus, the decision rule:

\[
y_i (w^T x_i + b) \geq 1 - \xi_i
\]

\[
\Rightarrow y_i (w^T x_i + b) \geq 0
\]

The point is on the correct side, but is on the margin and not sufficiently away.
Soft Margin SVM Classification

- Consider some cases for various $\xi_i$.

Let's consider point d that belongs to the other class: $y_i = -1$

Generally, for $y_i = -1$

$$y_i (w^T x_i + b) \geq 1 - \xi_i$$

$$\Rightarrow -(w^T x_i + b) \geq 1 - \xi_i$$

$$\Rightarrow w^T x_i + b < -1 + \xi_i$$

For point (d), since it is above the decision boundary, it should have $\xi_i > 1$

Hence, for misclassified points: $\xi_i > 1$.

$$y_i (\hat{w}^T \hat{x}_i + b) \geq 1 - \xi_i \quad \forall i$$
Soft Margin SVM Classification

- The value $\xi_i$ for **misclassified** points:
  
  c) Correctly classified, but inside the margin points: $\xi_i > 0$.
  
  d) Misclassified points: $\xi_i > 1$.

\[ y_i(\hat{w}^T \hat{x}_i + b) \geq 1 - \xi_i \quad \forall i \]
Soft Margin SVM Classification

- If $\xi_i = 0$, there is no problem with $x_i$.
- If $0 < \xi_i < 1$, $x_i$ is correctly classified but on the margin.
- If $\xi_i > 1$, $x_i$ is misclassified.

The number of misclassifications is $\#\{\xi_i > 1\}$.

The number of non-separable points is $\#\{\xi_i > 0\}$, e.g., “c”, “d”

$$y_i(\overrightarrow{w}^T \vec{x}_i + b) \geq 1 - \xi_i \quad \forall i$$
Soft Margin SVM Classification

• Thus, in the soft margin case our goal is two fold:
  - To maximize the margin (i.e., \( \min_{\vec{w}, b} \frac{1}{2} \vec{w}^T \vec{w} \))
  - Minimize non-separable points (\( \#\{ \xi_i > 0 \} \)).

• Based on this new consideration we update the minimization problem as follows.

\[
\begin{align*}
\min_{\vec{w}, b} & \frac{1}{2} ||\vec{w}||^2 \equiv \frac{1}{2} \vec{w}^T \vec{w} \\
\text{subject to } & y_i (\vec{w}^T \vec{x}_i + b) \geq 1 \forall i
\end{align*}
\]

\[
\begin{align*}
\min_{\vec{w}, b, \xi_i} & \frac{1}{2} \vec{w}^T \vec{w} + C \sum_{i=1}^{N} \xi_i \\
\text{subject to } & y_i (\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i \forall i \\
& \xi_i \geq 0
\end{align*}
\]
Soft Margin SVM Classification

- The new **formulation** of the **minimization problem** for the **soft margin** SVM classification includes the **slack variable** $\xi$.

\[
\min_{\mathbf{w}, b, \xi_i} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{N} \xi_i
\]

subject to \hspace{2cm} y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \hspace{2cm} \forall i

$\xi_i \geq 0$

- **C** is the **tradeoff parameter**
- **Smaller C**: cut more slack
- **Larger C**: No slack, be **strict to the violators**

Maximize the margin by **cutting some slack to the violators**

We can **control the amount of slack** we want to cut for the violators via “C”
Soft Margin SVM Classification

• Here, C is the **penalty factor**.
• It is **similar to the regularization** (**l_2 norm**) of the weight vector in linear and logistic regression.
• Note that we are penalizing not only the misclassified points but **also the ones in the margin for better generalization**.

\[
\begin{align*}
\min_{\vec{w}, b, \xi_i} & \quad \frac{1}{2} \vec{w}^T \vec{w} + C \sum_{i=1}^{N} \xi_i \\
\text{subject to} & \quad y_i (\vec{w}^T \hat{x}_i + b) \geq 1 - \xi_i \quad \forall i \\
\xi_i & \geq 0
\end{align*}
\]
SVM Soft-Margin Model: Constrained Optimization Algorithm
Soft Margin SVM: Primal Problem

• We have formulated the SVM soft margin classification as a **constrained optimization problem**.
• It is the primal form of the optimization problem.
• We can use a QP solver to find optimal w & b.
• However, the primal solution is inefficient for high-dimensional data.

\[
\min_{\vec{w}, b, \xi_i} \frac{1}{2} \vec{w}^T \vec{w} + C \sum_{i=1}^{N} \xi_i \\
\text{subject to } y_i (\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i \quad \forall i \\
\xi_i \geq 0
\]
Soft Margin SVM: Dual Problem

- To derive the **dual optimization** problem we need to transform the primal problem as an **unconstrained optimization problem**.

- Then we **analytically solve it** and plug its solution to get the dual version.

\[
\min_{\vec{w}, b, \xi_i} \frac{1}{2} \vec{w}^T \vec{w} + C \sum_{i=1}^{N} \xi_i \\
\text{subject to } y_i (\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i \quad \forall i \\
\xi_i \geq 0
\]
Soft Margin SVM: Dual Problem

• Let’s write the **Lagrangian** for the **soft margin case**.

\[
L(w^*, b^*, \alpha, \xi_i) = \frac{1}{2} w^{*T} w^* + C \sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i [1 - \xi_i - y_i((\bar{w})^T \bar{x}_i + b^*)] + \sum_{i=1}^{N} \mu_i(-\xi_i)
\]

Here \(\mu_i\)’s are the **new Lagrange multipliers** to guarantee the non-negativity of \(\xi_i\).
Soft Margin SVM: Dual Problem

• We take the **derivatives** with respect to the parameters and set them to 0:

\[
\frac{\partial}{\partial w} L(\vec{w}^*, b^*, \alpha, \xi_i) = w^T - \sum_{i=1}^{N} \alpha_i [y_i \vec{x}_i^T] = 0
\]

\[
\vec{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \vec{x}_i
\]

\[
\frac{\partial}{\partial b} L(\vec{w}^*, b^*, \alpha, \xi_i) = \sum_{i=1}^{N} \alpha_i [y_i] = 0
\]

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

\[
\frac{\partial}{\partial \alpha_i} L(\vec{w}^*, b^*, \alpha, \xi_i) = \sum_{i=1}^{N} \left[1 - \xi_i - y_i((\vec{w}^*)^T \vec{x}_i + b^*)\right] = 0
\]

\[
\frac{\partial}{\partial \xi_i} L(\vec{w}^*, b^*, \alpha, \xi_i) = C - \alpha_i - \mu_i = 0
\]

\[
L(\vec{w}^*, b^*, \alpha, \xi_i) = \frac{1}{2} \vec{w}^{*T} \vec{w}^* + C \sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i [1 - \xi_i - y_i((\vec{w}^*)^T \vec{x}_i + b^*)] - \sum_{i=1}^{N} \mu_i \xi_i
\]
Soft Margin SVM: Dual Problem

- Consider the last constraint: it’s a new constraint on $\alpha_i$.
- Since $\mu_i \geq 0$, the last constraint implies that $0 \leq \alpha_i \leq C$.

\[
\frac{\partial}{\partial \xi_i} \mathcal{L}(\vec{w}^*, b^*, \alpha, \xi_i) = C - \alpha_i - \mu_i = 0
\]

\[
\Rightarrow \alpha_i = C - \mu_i
\]

We derived **two constraints** on multiplier $\alpha_i$.

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

\[
0 \leq \alpha_i \leq C \quad \forall i
\]

\[
\mathcal{L}(\vec{w}^*, b^*, \alpha, \xi_i) = \frac{1}{2} \vec{w}^{*T} \vec{w}^* + C \sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i [1 - \xi_i - y_i((\vec{w}^*)^T \vec{x}_i + b^*)] - \sum_{i=1}^{N} \mu_i \xi_i
\]
Soft Margin SVM: Dual Problem

• Finally, we derive the dual for soft margin case (steps are similar to the hard margin case, hence not shown):

\[
\mathcal{L}(\vec{w}^*, b^*, \alpha, \xi_i) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j
\]

subject to \[
\sum_{i=1}^{N} \alpha_i^* y_i = 0
\]

\[
0 \leq \alpha_i \leq C \quad \forall i
\]

Find the optimal \(\alpha_i\) subject to the two constraints
Soft Margin SVM: Dual Problem

• Solving this, we see that as in the separable case, instances that lie on the correct side of the boundary with sufficient margin vanish with their $\alpha_i = 0$ (point at (a)).

• The support vectors have their $\alpha_i > 0$ and they define $\mathbf{w}$. 
Soft Margin SVM: Dual Problem

- Of these, those whose $\alpha_i < C$ are the ones that are on the margin, and we can use them to calculate $b$.
- They have $\xi_i = 0$ and satisfy $y_i(w^T x_i + b) = 1$.
- Again, it is better to take an average over these $b$ estimates.
- Those instances that are inside the margin or misclassified have their $\alpha_i = C$.

\[
0 \leq \alpha_i \leq C \quad \forall i
\]

\[
y_i(\overrightarrow{w}^T \overrightarrow{x}_i + b) \geq 1 - \xi_i \quad \forall i
\]
SVM Classifier: Error Analysis
Soft Margin SVM Classification

- **The nonseparable instances** that we store as **support vectors** (e.g., point c & d) are the instances that we would have trouble correctly classifying if they were not in the training set.
- They would **either be misclassified or classified correctly** but not with enough confidence.
Soft Margin SVM Classification

• We can say that the number of support vectors is an upper-bound estimate for the expected number of errors.

• Vapnik has shown that the expected test error rate is:

\[
E_N[P(error)] \leq \frac{E_N[\# \ of \ support \ vectors]}{N}
\]

Here \(E_N[\cdot]\) denotes expectation over training sets of size \(N\).

The nice implication of this is that it shows that the error rate depends on the number of support vectors and not on the input dimensionality.
Implementation
Issues
Soft Margin SVM Classification

• When implementing the SVM classifier, we need to pay attention to the following two issues.
  - Tuning the regularization/penalty parameter C
  - Scaling the features
Tuning the Regularization Parameter $C$
SVM Classification: Implementation Issue

- Note the in the SVM QP, **C is the regularization parameter.**
- It is fine-tuned using **cross-validation**.
- It defines the trade-off between **margin maximization and error minimization**.
  - If it is **too large**, we have a high penalty for nonseparable points, and we may store many support vectors and **overfit**.
  - If it is **too small**, we may find too simple solutions that **underfit**.

\[
\min_{\mathbf{w}, b, \xi_i} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{N} \xi_i
\]
SVM Classification: Implementation Issue

- On the left, using a high C value the classifier makes fewer margin violations but ends up with a smaller margin.
- On the right, using a low C value the margin is much larger, but many instances end up on the street.

\[
\min_{\vec{w}, b, \xi_i} \frac{1}{2} \vec{w}^T \vec{w} + C \sum_{i=1}^{N} \xi_i
\]
SVM Classification: Implementation

Issue

• However, it seems likely that the second classifier will generalize better.

• In fact even on this training set it makes fewer prediction errors, since most of the margin violations are actually on the correct side of the decision boundary.
Feature Scaling
SVM Classification: Implementation

Issue

• SVMs are sensitive to the feature scales.
• Consider the following scenario.

• **Left plot**: the vertical scale is much larger than the horizontal scale.

• So, the widest possible street is close to horizontal because SVM tends to ignore small scale features.

![Unscaled vs Scaled Plots](image_url)
SVM Classification: Implementation

Issue

• **Right plot**: the features are **scaled**.

• **After feature scaling**, the decision boundary looks much better (on the right plot).

Thus we must **scale the features** before training the SVM model.