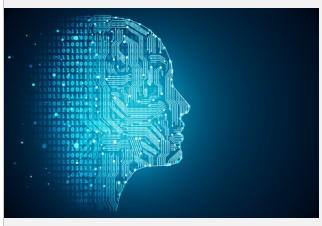


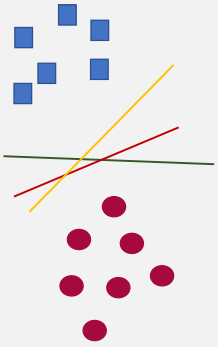
DS 310 Machine Learning
Maximum Margin Linear Classifiers

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

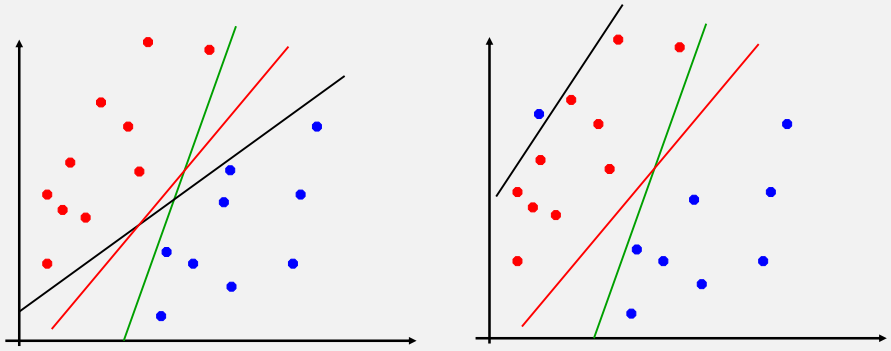
vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Linear classifiers revisited

- We have considered several algorithms that can find weight vectors that implement a linear hyperplane that correctly labels the training data when the data are separable.
- Perceptron algorithm finds a hyperplane (from among an infinite number of choices) that correctly classifies the training samples
- However, our goal is not just to correctly label the training samples, but to correctly predict the labels of data samples not observed in the training set
- Is one of the infinite number of hyperplanes optimal in the sense that it yields the lowest error on samples not seen in the training data (assuming they are also linearly separable)?



Which hyperplane to choose?



Linear classifiers differ with respect to:

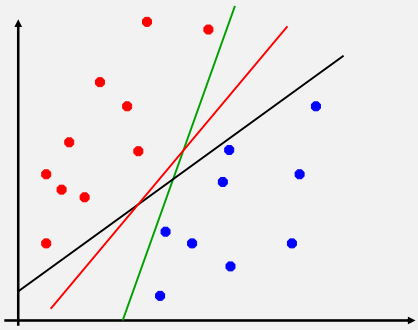
- which hyperplane they choose when data are linearly separable
- how they handle data that are not linearly separable

Which hyperplane to choose?

Perceptron:

- When data are separable, find some hyperplane that separates the data
- Could we do better?
- Is one of the infinite number of hyperplanes optimal in the sense that it yields the lowest error on samples not seen in the training data (assuming that the data are separable)?

Which hyperplane would you choose?



The Generalization Problem

- The Learning problem is ill posed
  - There are infinitely many hyperplanes that separate the training data
  - Need a principled approach to choose an optimal hyperplane

History of Key Developments

1958 Perceptron (Rosenblatt)

1963 Margin (Vapnik)

1964 Kernel Trick (Aizerman)

1965 Optimization formulation (Mangasarian)

1971 Kernels (Wahba)

1963-1992 SVM (Vapnik)

1996 – present Rapid growth, numerous applications
Extensions to other problems

Notation

In what follows, for consistency with the literature, we will use $\mathbf{w}$ to denote $[w_1 \cdots w_N]^T$ and b to denote w_0

and $\mathbf{x}$ to denote $[x_1 \cdots x_N]^T$

The linear hyperplane (or threshold function) is given by

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = 0$$

A Little Learning Theory

Suppose:

- We are given *l* training examples $(\mathbf{x}_i, d_i)$
- Train and test points drawn randomly (i.i.d) from some unknown probability distribution

- The machine learns the function $h_{\mathbf{w},b}(\mathbf{x}\quad)$
- A particular choice of $\mathbf{w}, b$ specifies "trained machine"
- The expectation $E(\mathbf{w}, b)$ of the test error is the probability that $h_{\mathbf{w},b}(\mathbf{x}\quad)$ incorrectly labels samples drawn according to the distribution that generated the training data
- The empirical estimate of $\hat{E}(\mathbf{w}, b) = \frac{1}{P}\sum_p^P \max\{-d_p h_{\mathbf{w},b}(\mathbf{x}_p), 0\}$

Bounding Generalization Error

Choose some δ such that $0 < \delta < 1$. With probability $1-\delta$ the following distribution independent bound holds (Vapnik,1995):

$$E(\mathbf{w}, b) \leq \hat{E}(\mathbf{w}, b) + \sqrt{\frac{d\left(\log\left(\frac{2P}{d}\right)+1\right)-\log\frac{\delta}{4}}{P}}$$

where $d \geq 0$ is called VC dimension is a measure of "capacity" of machine (proof omitted) and P is the number of training samples

- VC dimension is a key notion in theory of machine learning
- VC dimension impacts the number of samples needed to reliably learn a classifier of a certain type (e.g., a linear hyperplane)

Bounding Generalization Error

Choose some δ such that $0 < \delta < 1$. With probability $1-\delta$ the following distribution independent bound holds (Vapnik,1995):

$$E(\mathbf{w}, b) \leq \hat{E}(\mathbf{w}, b) + \sqrt{\frac{d\left(\log\left(\frac{2P}{d}\right)+1\right)-\log\frac{\delta}{4}}{P}}$$

where $d \geq 0$ is called VC dimension is a measure of "capacity" of machine

- If the training data are linearly separable, it is possible to achieve $\hat{E}(\mathbf{w}, b) = 0$.
- Then $E(\mathbf{w}, b)$= $\sqrt{\frac{d\left(\log\left(\frac{2P}{d}\right)+1\right)-\log\frac{\delta}{4}}{P}}$

  or $E(\mathbf{w}, b)$ can be minimized by controlling d and increasing P

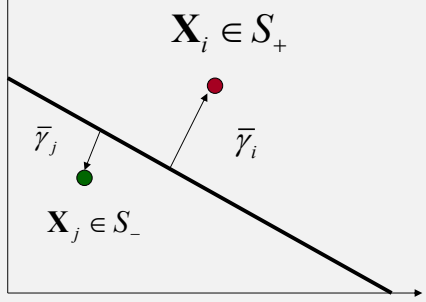
  Independent of the dimensionality of the input/feature space

Margin

- The functional margin of a linear hyperplane specified by $(\mathbf{w}, b)$ w.r.t. a labeled pattern $(\mathbf{x}_i, d_i)$ is defined as
$$\gamma_i = d_i\,(\mathbf{w} \cdot \mathbf{x}_i + b)$$
- If the functional margin is negative, then the pattern is incorrectly classified, if it is positive then the classifier predicts the correct label.
- The larger $|\gamma_i|$, the further away $\mathbf{x}_i$ is from the hyperplane
- This is made more precise in the notion of the geometric margin
$$\bar{\gamma}_i = \frac{\gamma_i}{\|\mathbf{w}\|}$$

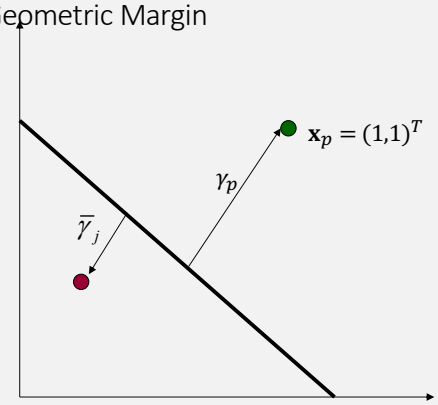
which measures the Euclidean distance of $\mathbf{x}_i$ from the decision boundary

Geometric Margin



The geometric margin of two points

Geometric Margin



Example

$$\mathbf{w} = (1,1)^T$$
$$b = -1$$
$$\mathbf{x}_p = (1,1)^T$$
$$d_p = 1$$
$$\gamma_p = (d_p)(\mathbf{w} \cdot \mathbf{x}_p \quad + b)$$
$$= (1)(1 + 1 - 1)$$
$$= 1$$
$$\bar{\gamma}_p = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{2}}$$

Margin of a hyperplane with respect to a data set

$$\overline{\gamma} = \min_p \frac{d_p\left(\mathbf{w} \cdot \mathbf{x}_p \quad + b\right)}{\|\mathbf{w}\|}$$

Margin based bound on generalization error

Margin based bound

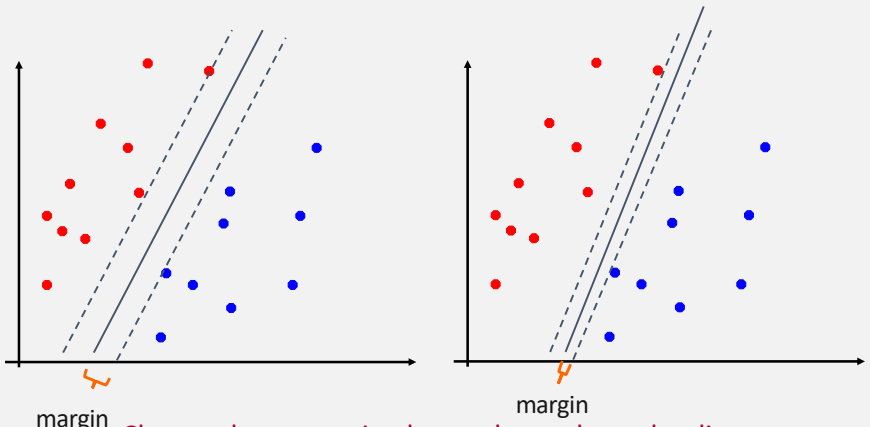
$$\bar{\gamma} = \min_i \frac{y_i f(\mathbf{x}_i)}{\|\mathbf{w}\|} \qquad \bar{\gamma} = \min_p \frac{d_p\left(\mathbf{w}\cdot\mathbf{x}_p + b\right)}{\|\mathbf{w}\|}$$

$$L = \max_p \|\mathbf{X}_p\| \qquad \varepsilon = O\left(\frac{1}{l}\left(\frac{L}{\bar{\gamma}}\right)^2\right)$$

Error $\mathcal{E}$ of the classifier trained on a separable data set is inversely proportional to its margin and is independent of the dimensionality of the input space! (proof omitted)

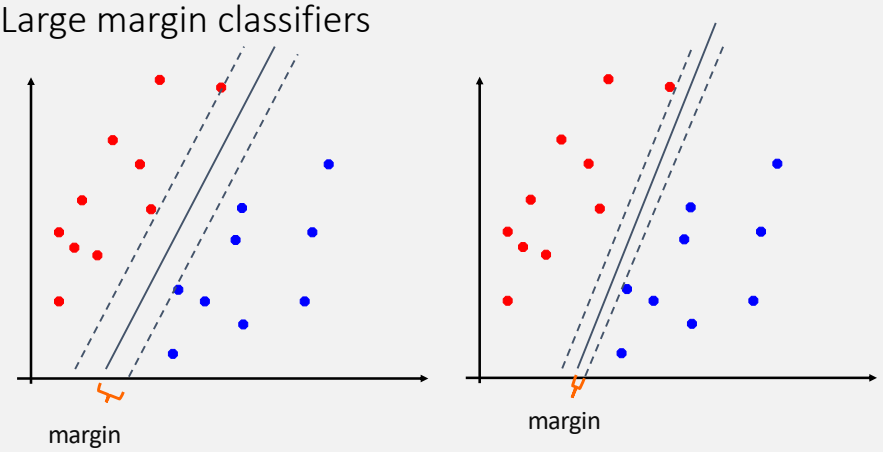
So if you want to minimize the error of the classifier on data unseen during training, we should pick a hyperplane with the largest margin on the training data!

Large margin classifiers



Choose the separating hyperplane where the distance to the nearest point(s) is as large as possible

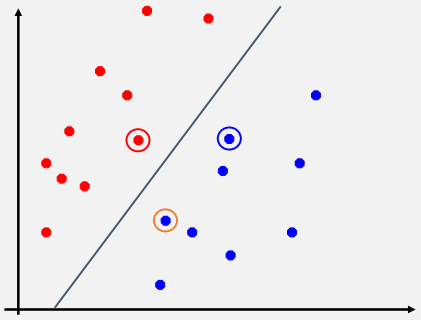
Large margin classifiers



- The margin of a classifier is the distance to the closest points of either class
- Large margin classifiers attempt to maximize this

Support vectors

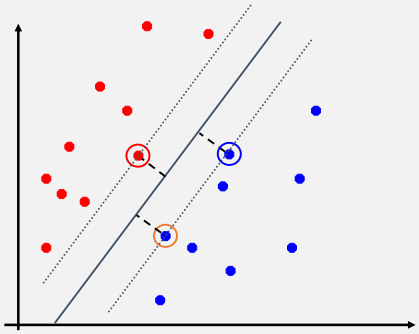
- For any separating hyperplane, there exist some set of "closest points"
- These are called the support vectors



- they are vectors (i.e. points) and they support/define the line

Measuring the margin

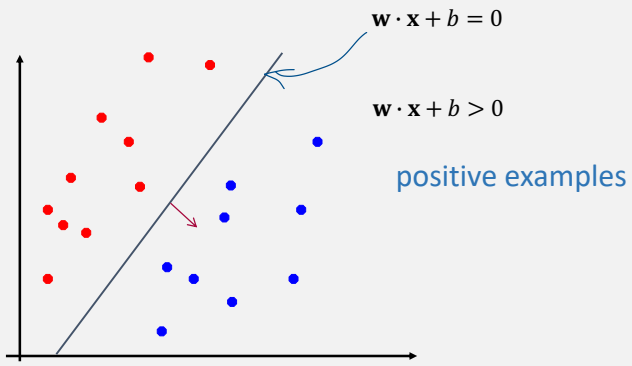
The margin of a hyperplane is the distance to the support vectors, i.e. the closest points, on either side of the hyperplane



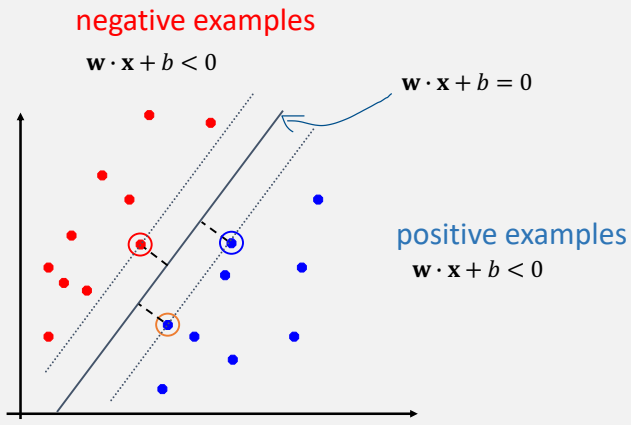
Measuring the margin

negative examples

$\mathbf{w} \cdot \mathbf{x} + b < 0$



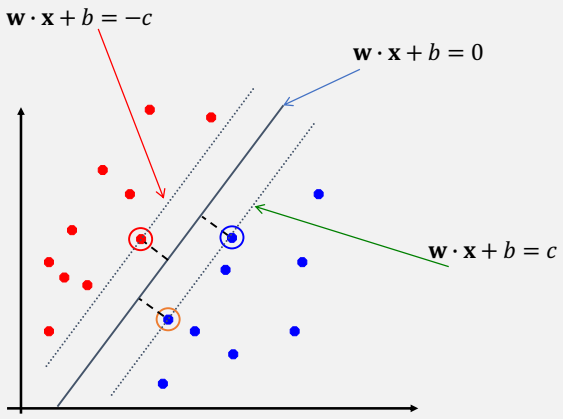
What defines the margin?



Maximum margin hyperplane

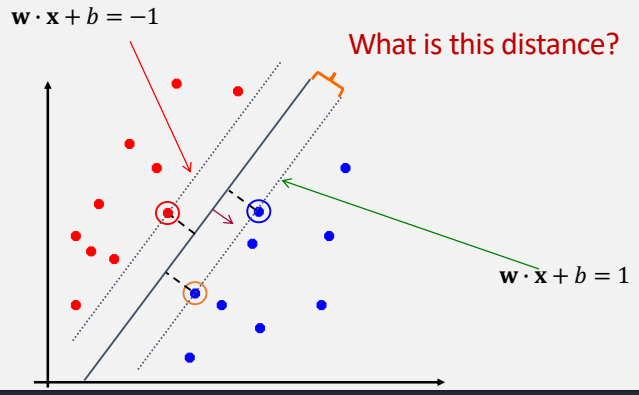
Place the hyperplane equidistant from the nearest points of the two classes

- What is c?
- It depends!
- If we scale $\mathbf{w}$, we can vary c without changing the separating hyperplane!



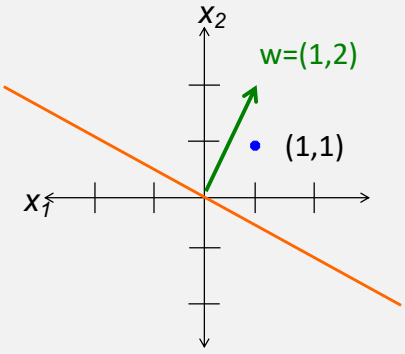
Maximum margin hyperplane

- If we scale $\mathbf{w}$, we can vary c without changing the separating hyperplane!
- So how about if we choose $c = 1$?



Distance from the hyperplane

How far away is the point (1,1) from the hyperplane?



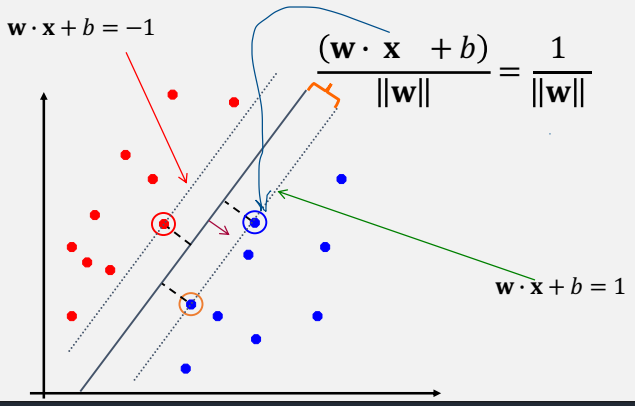
$$\frac{(\mathbf{w}\cdot\mathbf{x}_i + b)}{\|\mathbf{w}\|} = \frac{(1)(1)+(1)(2)+0}{\sqrt{1^2+2^2}}$$

$$= 1.34$$

- Note that the magnitude of the length of the weight vector does not matter!
- Only its direction and the value of b do.

Maximum margin hyperplane

- If we scale $\mathbf{w}$, we can vary c without changing the separating hyperplane!
- So how about if we choose $c = 1$?



Maximum margin classifier

Select, among all separating hyperplanes, i.e., one that correctly classifies the training samples with the largest geometric margin

This yields a constrained optimization problem:

$$\max_{\mathbf{w},b} margin(\mathbf{w}, b)$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$$

Because $margin(\mathbf{w}, b) = \frac{1}{\|\mathbf{w}\|}$, maximizing the margin is equivalent to minimizing $\|\mathbf{w}\|$ subject to the constraints $\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$

All points are classified correctly AND they have a prediction >= 1.

Maximum margin classifier

Select, among all separating hyperplanes, i.e., one that correctly classifies the training samples with the largest geometric margin

- Because $margin(\mathbf{w}, b) = \frac{1}{\|\mathbf{w}\|}$, maximizing the margin is equivalent to minimizing $\|\mathbf{w}\|$ subject to the constraints $\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$
- This yields the following constrained optimization problem:

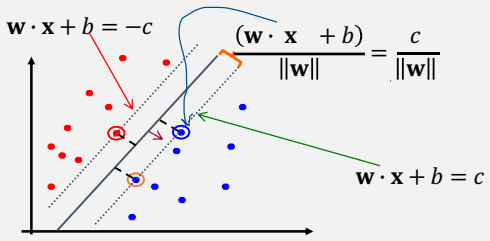
$$\min_{\mathbf{w},b} \|\mathbf{w}\|$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$$

All points are classified correctly AND they have a prediction >= 1.

Maximum margin hyperplane



- If we scale $\mathbf{w}$, we can vary c without changing the separating hyperplane!

$$\min_{\mathbf{w},b}\frac{\|\mathbf{w}\|}{c}$$

Subject to:

$$\forall p \quad d_p\left(\mathbf{w}\cdot\mathbf{x}_p+b\right)\geq c$$

Maximum margin hyperplane

Is there a difference between

$$\min_{\mathbf{w},b} \|\mathbf{w}\|$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$$

and

$$\min_{\mathbf{w},b} \frac{\|\mathbf{w}\|}{c}$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq c$$

- No!
- They are both equivalent, modulo the scaling of $\|\mathbf{w}\|$

Maximum margin hyperplane

We have to

$$\min_{\mathbf{w},b} \|\mathbf{w}\|$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$$

We consider instead

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$$

Why?

- $\min_{\mathbf{w},b} \|\mathbf{w}\|^2$ implies $\min_{\mathbf{w},b} \|\mathbf{w}\|$
- $\|\mathbf{w}\|^2$ is a convex function of $\mathbf{w}$!

Maximum margin classifier – Support vector machine

Quadratic optimization problem

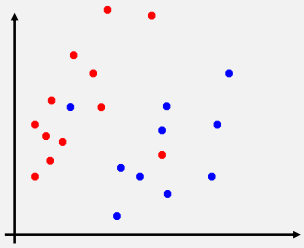
$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$$

- Note that we have to restrict the domain of $\mathbf{w}$ over which $\|\mathbf{w}\|^2$ is minimized to the region where the constraints $\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$ are satisfied
- There are many algorithms to solve such constrained optimization problems
- We will see one shortly

Soft Margin Classification



- What if the data are not linearly separable?
- There is no solution to

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$$

- What can we do?
  - Allow soft margins instead of hard margins
  - That is, allow some samples to be misclassified

Soft Margin Classification: Slack variables to the rescue

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1$$

Introduce slack variables, one for each sample:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

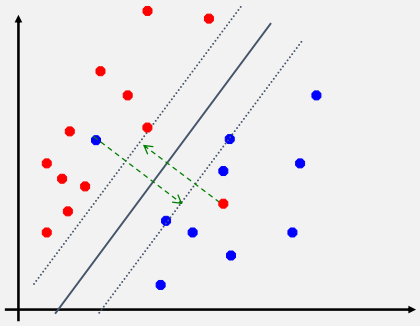
Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1 - \zeta_p$$

and $\zeta_p \geq 0$

What do the slack variables do?

Slack variables



$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1 - \zeta_p$$

and $\zeta_p \geq 0$

slack variables penalize misclassified samples

How do slack variables help?

Maximize margin

trade-off between margin maximization and

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

Penalize by the distance wrong side of the hyperplane

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right) \geq 1 - \zeta_p$$

and $\zeta_p \geq 0$

Allow an occasional misclassification

Soft margin Support Vector Machine

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

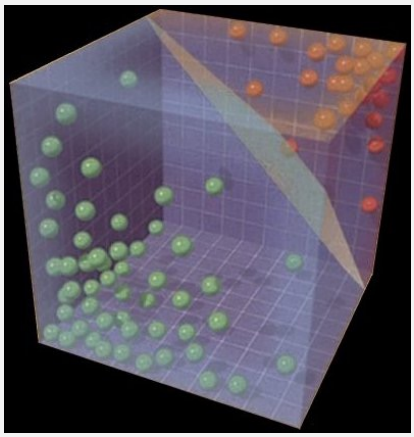
Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right) \geq 1 - \zeta_p$$

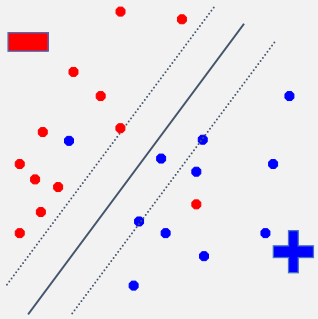
and $\zeta_p \geq 0$

- Still a quadratic optimization problem!
- Quadratic objective function and linear constraints

Solving the soft margin SVM



Understanding the Soft Margin SVM



$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right) \geq 1 - \zeta_p$$

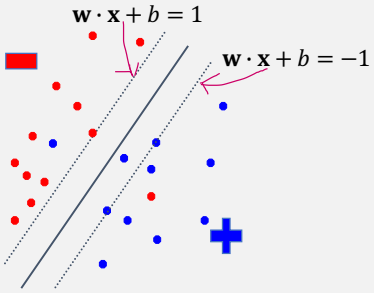
and $\zeta_p \geq 0$

Given the optimal solution, $\mathbf{w}, b$:

- Can we figure out what the slack penalties are for each sample?

Understanding the Soft Margin SVM

What do the margin lines represent w.r.t $\mathbf{w}, b$?



$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

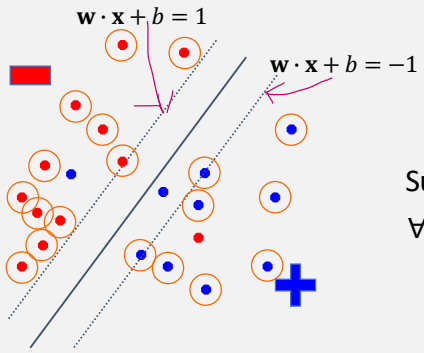
Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right) \geq 1 - \zeta_p$$

and $\zeta_p \geq 0$

Thus, at the margins $d_p(\mathbf{w} \cdot \mathbf{x} + b) = 1$

Understanding the Soft Margin SVM



$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

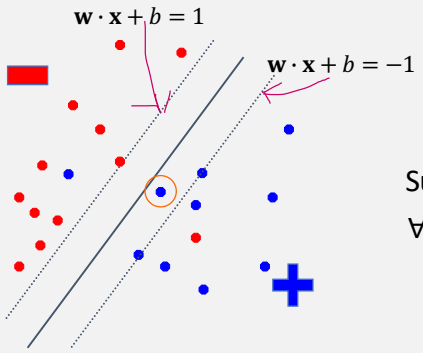
Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right) \geq 1 - \zeta_p$$

and $\zeta_p \geq 0$

- What are the slack values for correctly classified samples that are on or outside the margin?
  - $d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right) \geq 1$, so it must be the case that $\zeta_p = 0$

Understanding the Soft Margin SVM



$$\min_{\mathbf{w},b}\|\mathbf{w}\|^2 + C\sum_p \zeta_p$$

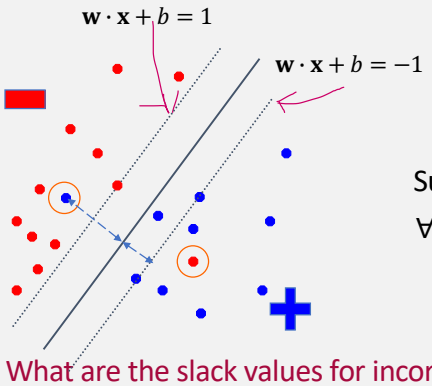
Subject to:

$$\forall p \quad d_p\left(\mathbf{w}\cdot\mathbf{x}_p + b\right) \geq 1-\zeta_p$$

and $\zeta_p \geq 0$

- What are the slack values for correctly classified samples that are inside the margin?
  - $\zeta_p = 1 - d_p\left(\mathbf{w}\cdot\mathbf{x}_p + b\right)$

Understanding the Soft Margin SVM



$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right) \geq 1 - \zeta_p$$

and $\zeta_p \geq 0$

- What are the slack values for incorrectly classified samples that are on the wrong side of the hyperplane?
  - The distance to the hyperplane + distance to the margin
  - $\zeta_p = 1 - d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right)$

Understanding the Soft Margin SVM

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1 - \zeta_p$$

and $\zeta_p \geq 0$

$$\zeta_p = \begin{cases} 0 \text{ if } d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1 \\ 1 - d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \text{ otherwise} \end{cases}$$

In other words

$$\zeta_p = \max \left\{ 0, \left(1 - d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \right) \right\}$$

Understanding the Soft Margin SVM

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

Subject to:

$$\forall p \quad d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \geq 1 - \zeta_p$$

and $\zeta_p \geq 0$

The above implies

$$\zeta_p = \max \left\{ 0, \left(1 - d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b \right) \right) \right\}$$

Do we still need the constraints?

Not really, because we can substitute ζ_p into the objective function!

Understanding the Soft Margin SVM

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \zeta_p$$

is equivalent to

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_p \max\left\{0, \left(1 - d_p \left(\mathbf{w} \cdot \mathbf{x}_p \quad + b\right)\right)\right\}$$

The result is an unconstrained optimization problem

Solving the Soft Margin SVM

We want to find $\mathbf{w}, b$ that minimize

$$E(\mathbf{w}, b) = \|\mathbf{w}\|^2 + C \sum_p \max\left\{0, \left(1 - d_p\left(\mathbf{w} \cdot \mathbf{x}_p + b\right)\right)\right\}$$

We note that this objective function is convex and is differentiable everywhere except where $1 - d_p\left(\mathbf{w} \cdot \mathbf{x}_p + b\right) = 0$

So we can find the (sub) gradients with respect to $\mathbf{w}$, and b to and iteratively update them in the direction of their negative gradients

Solving the Soft Margin SVM

We want to find $\mathbf{w}, b$ that minimize

$$E(\mathbf{w}, b) = \|\mathbf{w}\|^2 + C \sum_p \max\left\{0, \left(1 - d_p\left(\mathbf{w} \cdot \mathbf{x}_p + b\right)\right)\right\}$$

- When $d_p\left(\mathbf{w} \cdot \mathbf{x}_p + b\right) \geq 1$
  - the second term in $E(\mathbf{w}, b)$ is 0
  - $\frac{\partial E(\mathbf{w},b)}{\partial \mathbf{w}} = 2\,\mathbf{w}$ and $\frac{\partial E(\mathbf{w},b)}{\partial b} = 0$
- When $d_p\left(\mathbf{w} \cdot \mathbf{x}_p + b\right) < 1$
  - the second term in $E(\mathbf{w}, b)$ is $C \sum_p \left(1 - d_p\left(\mathbf{w} \cdot \mathbf{x}_p + b\right)\right)$
  - $\frac{\partial E(\mathbf{w},b)}{\partial \mathbf{w}} = 2\,\mathbf{w} - C \sum_p d_p \mathbf{x}_p$ and $\frac{\partial E(\mathbf{w},b)}{\partial b} = -C \sum_p d_p$

Solving the Soft Margin SVM

Combining the two cases, the weight update rule is

- $\mathbf{w} \leftarrow \mathbf{w} + \eta \left(\mathrm{I} \left(d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right) < 1 \right) C \sum_p d_p \mathbf{x}_p - 2\mathbf{w} \right)$
- $b \leftarrow b + \eta \mathrm{I} \left(d_p \left(\mathbf{w} \cdot \mathbf{x}_p + b \right) < 1 \right) C \sum_p d_p$

Where $\mathrm{I}(z)$ is the indicator function. $\mathrm{I}(z) = 1$ if z is True, and 0 otherwise.

- C is tuned using cross-validation.

Remarks on the soft margin SVM

- Soft margin SVM finds the maximum margin classifier if the data are linearly separable
- Theory tells us that the maximum margin classifier will generalize optimally on unseen samples if the classes are linearly separable
- If the data are not linearly separable, soft margin SVM produces a classifier implements a good compromise
- We used the hinge loss $\max\left\{0, \left(1 - d_p\left(\mathbf{w} \cdot \mathbf{x}_p \quad + b\right)\right)\right\}$ in our objective function
- We can replace it by a smooth differentiable approximation as we saw in the case of the perceptron and derive the update equations. This is left as an exercise.
- Later we will generalize SVMs to incorporate kernel functions so they can be applied to data that do not necessarily live in a Euclidian space and in settings where the data may not be linearly separable