# Support Vector Machines 

## Nicholas Ruozzi <br> University of Texas at Dallas

## Announcements

- Homework 1 available soon
- Piazza - join if you haven't already
- Reminder: my office hours are 10am-11am on Tuesdays in ECSS 3.409
- Schedule and lecture notes available through the course website (see the link on eLearning)


## Binary Classification

- Input $\left(x^{(1)}, y^{(1)}\right), \ldots,\left(x^{(M)}, y^{(M)}\right)$ with $x^{(m)} \in \mathbb{R}^{n}$ and $y^{(m)} \in$ $\{-1,+1\}$
- We can think of the observations as points in $\mathbb{R}^{n}$ with an associated sign (either + - corresponding to 0/1)



## Binary Classification

- $\operatorname{Input}\left(x^{(1)}, y^{(1)}\right), \ldots,\left(x^{(M)}, y^{(M)}\right)$ with $x^{(m)} \in \mathbb{R}^{n}$ and $y^{(m)} \in$ $\{-1,+1\}$
- We can think of the observations as points in $\mathbb{R}^{n}$ with an associated sign (either $+/$ - corresponding to $0 / 1$ )



## What If the Data Isn't Separable?

- $\operatorname{Input}\left(x^{(1)}, y^{(1)}\right), \ldots,\left(x^{(M)}, y^{(M)}\right)$ with $x^{(m)} \in \mathbb{R}^{n}$ and $y^{(m)} \in$ $\{-1,+1\}$
- We can think of the observations as points in $\mathbb{R}^{n}$ with an associated sign (either + /- corresponding to 0/1)

What is a good
hypothesis space for
this problem?

## What If the Data Isn't Separable?

- $\operatorname{Input}\left(x^{(1)}, y^{(1)}\right), \ldots,\left(x^{(M)}, y^{(M)}\right)$ with $x^{(m)} \in \mathbb{R}^{n}$ and $y^{(m)} \in$ $\{-1,+1\}$
- We can think of the observations as points in $\mathbb{R}^{n}$ with an associated sign (either + /- corresponding to 0/1)


What is a good hypothesis space for this problem?

## Adding Features

- The idea:
- Given the observations $x^{(1)}, \ldots, x^{(M)}$, construct a feature vectors $\phi\left(x^{(1)}\right), \ldots, \phi\left(x^{(M)}\right)$
- Use $\phi\left(x^{(1)}\right), \ldots, \phi\left(x^{(M)}\right)$ instead of $x^{(1)}, \ldots, x^{(M)}$ in the learning algorithm
- Goal is to choose $\phi$ so that $\phi\left(x^{(1)}\right), \ldots, \phi\left(x^{(M)}\right)$ are linearly separable
- Learn linear separators of the form $w^{T} \phi(x)\left(\right.$ instead of $\left.w^{T} x\right)$
- Warning: more expressive features can lead to overfitting!


## Support Vector Machines

- How can we decide between perfect classifiers?



## Support Vector Machines

- How can we decide between perfect classifiers?



## Support Vector Machines

- Define the margin to be the distance of the closest data point to the classifier



## Support Vector Machines

- Support vector machines (SVMs)

- Choose the classifier with the largest margin
- Has good practical and theoretical performance


## Some Geometry



- In $n$ dimensions, a hyperplane is a solution to the equation

$$
w^{T} x+b=0
$$

with $w \in \mathbb{R}^{n}, b \in \mathbb{R}$

- The vector $w$ is sometimes called the normal vector of the hyperplane


## Some Geometry



- In $n$ dimensions, a hyperplane is a solution to the equation

$$
w^{T} x+b=0
$$

- Note that this equation is scale invariant for any scalar $c$

$$
c \cdot\left(w^{T} x+b\right)=0
$$

## Some Geometry



- The distance between a point $y$ and a hyperplane $w^{T}+b=0$ is the length of the segment perpendicular to the line to the point $y$
- The vector from $y$ to $z$ is given by

$$
y-z=\|y-z\| \frac{w}{\|w\|}
$$

## Scale Invariance



- By scale invariance, we can assume that $c=1$
- The maximum margin is always attained by choosing $w^{T} x+$ $b=0$ so that it is equidistant from the closest data point classified as +1 and the closest data point classified as -1


## Scale Invariance



- We want to maximize the margin subject to the constraints that

$$
y^{(i)}\left(w^{T} x^{(i)}+b\right) \geq 1
$$

- But how do we compute the size of the margin?


## Some Geometry



Putting it all together

$$
y-z=\|y-z\| \frac{w}{\|w\|}
$$

and

$$
\begin{aligned}
& w^{T} y+b=1 \\
& w^{T} z+b=0
\end{aligned}
$$

$$
w^{T}(y-z)=1
$$

and

$$
w^{T}(y-z)=\|y-z\|\|w\|
$$

which gives

$$
\|y-z\|=1 /\|w\|
$$

## SVMs

- This analysis yields the following optimization problem

$$
\max _{w} \frac{1}{\|w\|}
$$

such that

$$
y^{(i)}\left(w^{T} x^{(i)}+b\right) \geq 1, \text { for all } i
$$

- Or, equivalently,

$$
\min _{w}\|w\|^{2}
$$

such that

$$
y^{(i)}\left(w^{T} x^{(i)}+b\right) \geq 1, \text { for all } i
$$

## SVMs

$$
\min _{w}\|w\|^{2}
$$

such that

$$
y^{(i)}\left(w^{T} x^{(i)}+b\right) \geq 1, \text { for all } i
$$

- This is a standard quadratic programming problem
- Falls into the class of convex optimization problems
- Can be solved with many specialized optimization tools (e.g., quadprog() in MATLAB)


## SVMs



- Where does the name come from?
- The set of all data points such that $y^{(i)}\left(w^{T} x^{(i)}+b\right)=1$ are called support vectors


## SVMs

- What if the data isn't linearly separable?
- Use feature vectors
- Relax the constraints (coming soon)
- What if we want to do more than just binary classification (i.e., if $y \in\{1,2,3\}$ )?


## Multiclass Classification



## One-Versus-All SVMs



## One-Versus-All SVMs



Regions correctly classified by exactly one classifier

## One-Versus-All SVMs

- Compute a classifier for each label versus the remaining labels (i.e., and SVM with the selected label as plus and the remaining labels changed to minuses)
- Let $f^{k}(x)=w^{(k)^{T}} x+b^{(k)}$ be the classifier for the $k^{t h}$ label
- For a new datapoint $x$, classify it as

$$
k^{\prime} \in \operatorname{argmax}_{k} f^{k}(x)
$$

- Drawbacks:
- If there are $L$ possible labels, requires learning $L$ classifiers over the entire data set
- Doesn't make sense if the classifiers are not comparable


## One-Versus-All SVMs



Regions in which points are classified by highest value of $w^{T} x+b$

## One-Versus-One SVMs

- Alternative strategy is to construct a classifier for all possible pairs of labels
- Given a new data point, can classify it by majority vote (i.e., find the most common label among all of the possible classifiers)
- If there are $L$ labels, requires computing $\binom{L}{2}$ different classifiers each of which uses only a fraction of the data
- Drawbacks: Can overfit if some pairs of labels do not have a significant amount of data (plus it can be computationally expensive)


## One-Versus-One SVMs



Regions determined by majority vote over the classifiers

