Machine Learning

A. G. Schwing & M. Telgarsky

University of Illinois at Urbana-Champaign, 2018

L7: Multiclass Classification and Kernel Methods

Announcements:

- Office hours on Friday 5-7pm (2TAs) ECEB 2015
- Office hours Instructors: right after this class; questions about any of the material

Goals of this lecture

Goals of this lecture

- Understanding multi-class logistic regression

Goals of this lecture

- Understanding multi-class logistic regression
- Getting to know multi-class SVM

Goals of this lecture

- Understanding multi-class logistic regression
- Getting to know multi-class SVM
- Learning the kernel trick

Goals of this lecture

- Understanding multi-class logistic regression
- Getting to know multi-class SVM
- Learning the kernel trick

Reading material

Goals of this lecture

- Understanding multi-class logistic regression
- Getting to know multi-class SVM
- Learning the kernel trick

Reading material

- K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 14

Recap:

Recap:

- Linear regression:

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T \phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T \phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

- Logistic regression:

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

- Logistic regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})\right)$$

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

- Logistic regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})\right)$$

- Binary SVM:

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

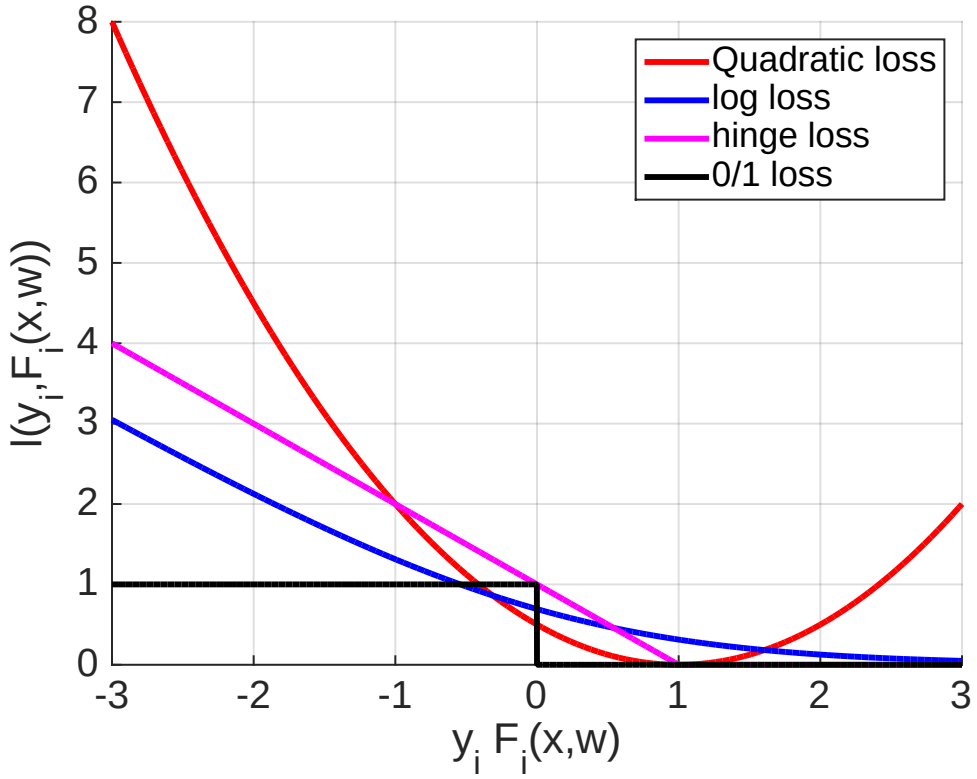
- Logistic regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})\right)$$

- Binary SVM:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \max\{0, \underbrace{1}_{\text{taskloss}} - y^{(i)} \underbrace{\boldsymbol{w}^\top\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})}\}$$

Recap:


Quadratic loss
log loss
hinge loss
0/1 loss
8
7
6
5
4
3
2
1
0
-3
-2
-1
0
1
2
3
l(y_i,F_i(x,w))
y_i F_i(x,w)

Other loss functions:

Other loss functions:

- Generalization of log- and hinge-loss

Other loss functions:

- Generalization of log- and hinge-loss
- Ramp loss minimization

Other loss functions:

- Generalization of log- and hinge-loss
- Ramp loss minimization
- Orbit loss minimization

Other loss functions:

- Generalization of log- and hinge-loss
- Ramp loss minimization
- Orbit loss minimization
- Direct loss minimization

Combining log- and hinge-loss:

$$\lim_{\epsilon \to 0} \epsilon \log \left(1 + \exp \frac{-F}{\epsilon} \right) = \underset{\equiv}{F \geq 0}$$

Combining log- and hinge-loss:

$$\lim_{\epsilon \to 0} \epsilon \log \left(1 + \exp \frac{-F}{\epsilon} \right) = \overset{F \geq 0}{\equiv} 0$$

Combining log- and hinge-loss:

$$\lim_{\epsilon \to 0} \epsilon \log \left(1 + \exp \frac{-F}{\epsilon} \right) =$$

$$\overset{F \geq 0}{=} 0$$

$$\overset{F \leq 0}{=}$$

Combining log- and hinge-loss:

$$\lim_{\epsilon \to 0} \epsilon \log \left(1 + \exp \frac{-F}{\epsilon}\right) =$$

$$\overset{F \geq 0}{=} 0$$

$$\overset{F \leq 0}{=} \lim_{\epsilon \to 0} \frac{\log \left(1 + \exp \frac{-F}{\epsilon}\right)}{1/\epsilon}$$

Combining log- and hinge-loss:

$$\lim_{\epsilon \to 0} \epsilon \log \left(1 + \exp \frac{-F}{\epsilon}\right) =$$

$$\overset{F \geq 0}{=} 0$$

$$\overset{F \leq 0}{=} \lim_{\epsilon \to 0} \frac{\log \left(1 + \exp \frac{-F}{\epsilon}\right)}{1/\epsilon}$$

$$= \lim_{\epsilon \to 0} \frac{\frac{\exp \frac{-F}{\epsilon}}{1+\exp \frac{-F}{\epsilon}} \cdot (F/\epsilon^2)}{-1/\epsilon^2}$$

Combining log- and hinge-loss:

$$\lim_{\epsilon \to 0} \epsilon \log\left(1 + \exp \frac{-F}{\epsilon}\right) =$$

$$\overset{F \geq 0}{=} 0$$

$$\overset{F \leq 0}{=} \lim_{\epsilon \to 0} \frac{\log\left(1 + \exp \frac{-F}{\epsilon}\right)}{1/\epsilon}$$

$$= \lim_{\epsilon \to 0} \frac{\frac{\exp \frac{-F}{\epsilon}}{1+\exp \frac{-F}{\epsilon}} \cdot (F/\epsilon^2)}{-1/\epsilon^2}$$

$$= \lim_{\epsilon \to 0} \frac{1}{1 + \exp \frac{F}{\epsilon}} \cdot (-F)$$

Combining log- and hinge-loss:

$$\lim_{\epsilon \to 0} \epsilon \log \left(1 + \exp \frac{-F}{\epsilon}\right) =$$

$$\overset{F \geq 0}{=} 0$$

$$\overset{F \leq 0}{=} \lim_{\epsilon \to 0} \frac{\log \left(1 + \exp \frac{-F}{\epsilon}\right)}{1/\epsilon}$$

$$= \lim_{\epsilon \to 0} \frac{\frac{\exp \frac{-F}{\epsilon}}{1+\exp \frac{-F}{\epsilon}} \cdot (F/\epsilon^2)}{-1/\epsilon^2}$$

$$= \lim_{\epsilon \to 0} \frac{1}{1 + \exp \frac{F}{\epsilon}} \cdot (-F)$$

$$= -F$$

Combining log- and hinge-loss:

$$\lim_{\epsilon \to 0} \epsilon \log\left(1 + \exp \frac{-F}{\epsilon}\right) =$$

$$\overset{F \geq 0}{=} 0$$

$$\overset{F \leq 0}{=} \lim_{\epsilon \to 0} \frac{\log\left(1 + \exp \frac{-F}{\epsilon}\right)}{1/\epsilon}$$

$$= \lim_{\epsilon \to 0} \frac{\frac{\exp \frac{-F}{\epsilon}}{1+\exp \frac{-F}{\epsilon}} \cdot (F/\epsilon^2)}{-1/\epsilon^2}$$

$$= \lim_{\epsilon \to 0} \frac{1}{1 + \exp \frac{F}{\epsilon}} \cdot (-F)$$

$$= -F$$

In summary:

$$\lim_{\epsilon \to 0} \epsilon \log\left(1 + \exp \frac{-F}{\epsilon}\right) = \max\{0, -F\}$$

Combining log- and hinge-loss:

$$\lim_{\epsilon \to 0} \epsilon \log\left(1 + \exp \frac{-F}{\epsilon}\right) =$$

$$\overset{F \geq 0}{=} 0$$

$$\overset{F \leq 0}{=} \lim_{\epsilon \to 0} \frac{\log\left(1 + \exp \frac{-F}{\epsilon}\right)}{1/\epsilon}$$

$$= \lim_{\epsilon \to 0} \frac{\frac{\exp \frac{-F}{\epsilon}}{1+\exp \frac{-F}{\epsilon}} \cdot (F/\epsilon^2)}{-1/\epsilon^2}$$

$$= \lim_{\epsilon \to 0} \frac{1}{1 + \exp \frac{F}{\epsilon}} \cdot (-F)$$

$$= -F$$

In summary:

$$\lim_{\epsilon \to 0} \epsilon \log\left(1 + \exp \frac{-F}{\epsilon}\right) = \max\{0, -F\}$$

SVM as 0-temperature limit of logistic regression

Recap:

Recap:

- Linear regression:

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T \phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T \phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

- Logistic regression:

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

- Logistic regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})\right)$$

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

- Logistic regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})\right)$$

- Binary SVM:

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

- Logistic regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})\right)$$

- Binary SVM:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \max\{0, \underbrace{1}_{\text{taskloss}} - y^{(i)} \underbrace{\boldsymbol{w}^\top\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})}\}$$

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

- Logistic regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})\right)$$

- Binary SVM:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \max\{0, \underbrace{1}_{\text{taskloss}} - y^{(i)} \underbrace{\boldsymbol{w}^\top\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})}\}$$

- General binary classification:

Recap:

- Linear regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \frac{1}{2}(1 - y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})^2$$

- Logistic regression:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)} \underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})})\right)$$

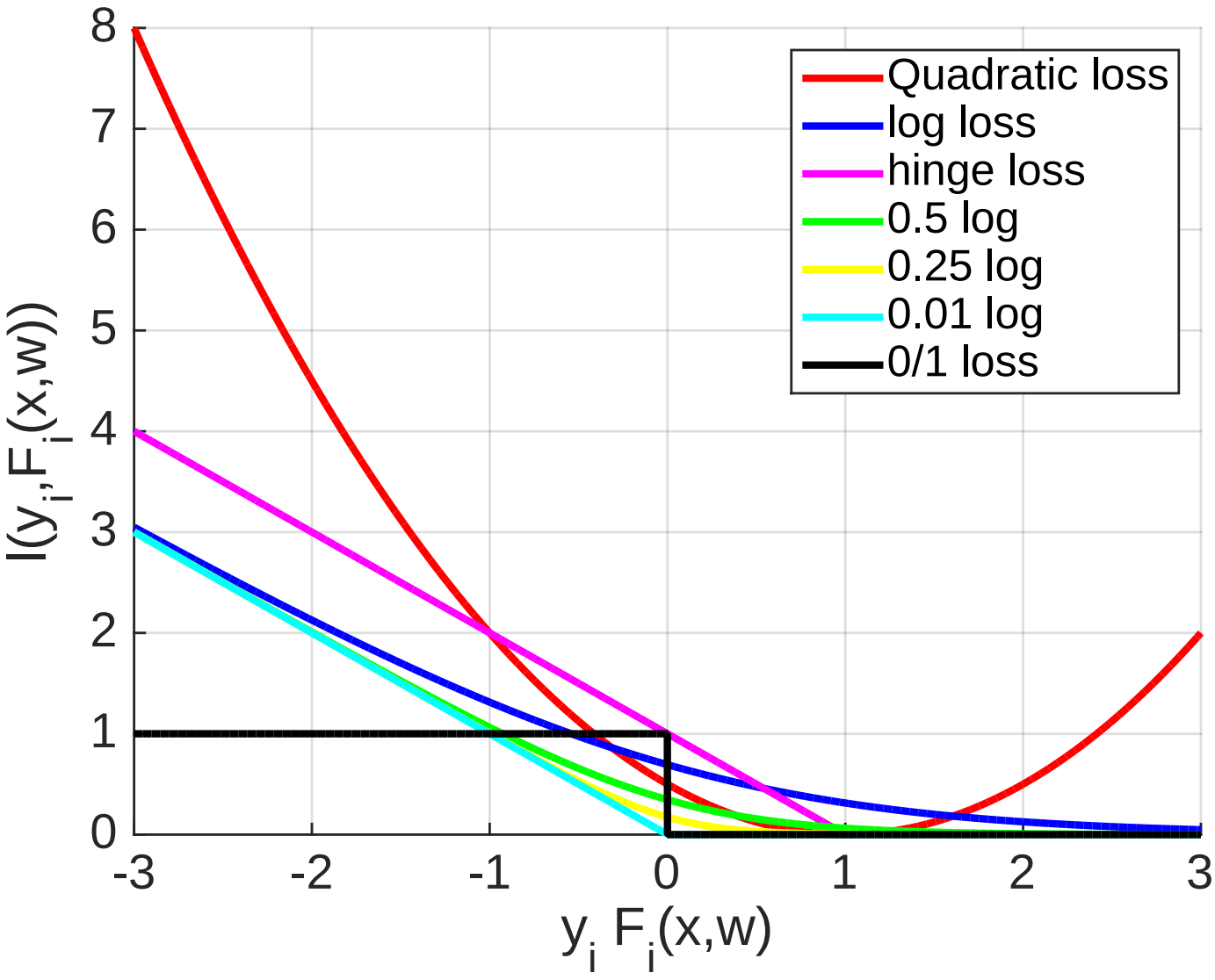
- Binary SVM:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \max\{0, \underbrace{1}_{\text{taskloss}} - y^{(i)} \underbrace{\boldsymbol{w}^\top\phi(x^{(i)})}_{F(x^{(i)},\boldsymbol{w})}\}$$

- General binary classification:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \epsilon \log\left(1 + \exp\left(\frac{L - y^{(i)}\boldsymbol{w}^T\phi(x^{(i)})}{\epsilon}\right)\right)$$

Different loss functions



Tasks we considered so far:

Tasks we considered so far:

- Edge detection

Tasks we considered so far:

- Edge detection
- Object detection

Tasks we considered so far:

- Edge detection
- Object detection

Common Theme between those task:

Tasks we considered so far:

- Edge detection
- Object detection

Common Theme between those task:

- Is a given pixel at a boundary?

Tasks we considered so far:

- Edge detection
- Object detection

Common Theme between those task:

- Is a given pixel at a boundary? **Yes/No**

Tasks we considered so far:

- Edge detection
- Object detection

Common Theme between those task:

- Is a given pixel at a boundary? **Yes/No**
- Is there a cat in the image?

Tasks we considered so far:

- Edge detection
- Object detection

Common Theme between those task:

- Is a given pixel at a boundary? **Yes/No**
- Is there a cat in the image? **Yes/No**

Image classification: What do we want?

Image classification: What do we want?

Which object is illustrated?

Image classification: What do we want?

Which object is illustrated?

- Car

Image classification: What do we want?

Which object is illustrated?

- Car
- Truck

Image classification: What do we want?

Which object is illustrated?

- Car
- Truck
- Recreational Vehicle

Image classification: What do we want?

Which object is illustrated?

- Car
- Truck
- Recreational Vehicle
- Ambulance truck

Image classification: What do we want?

Which object is illustrated?

- Car
- Truck
- Recreational Vehicle
- Ambulance truck
- Fire truck

Image classification: What do we want?

Which object is illustrated?

- Car
- Truck
- Recreational Vehicle
- Ambulance truck
- Fire truck

What's the difference to the tasks we considered so far?

Image classification: What do we want?

Which object is illustrated?

- Car
- Truck
- Recreational Vehicle
- Ambulance truck
- Fire truck

What's the difference to the tasks we considered so far?

More than two answers (multiple classes) need to be considered.

Multiclass classification: How to classify between K classes?

Multiclass classification: How to classify between K classes?

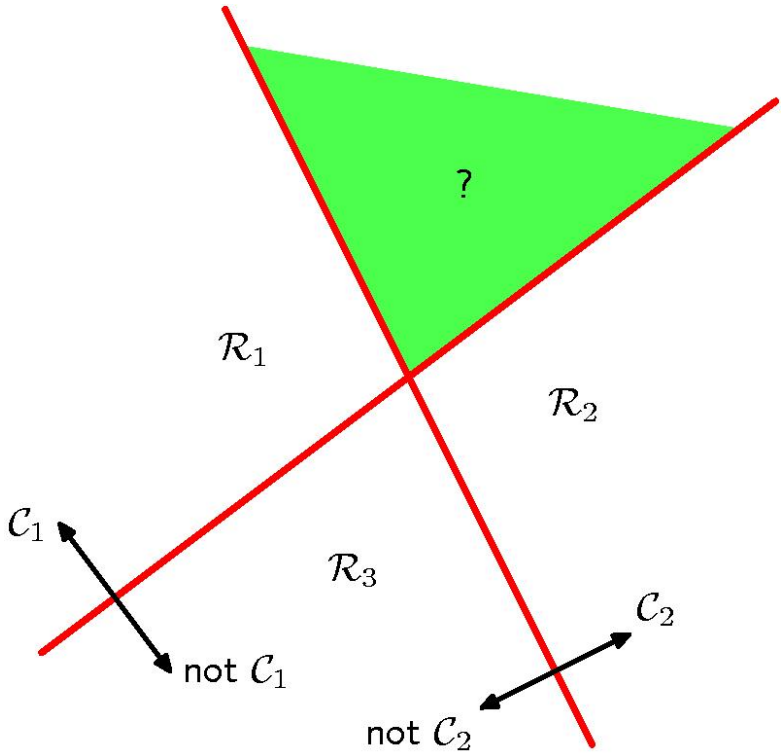
- **1 vs all** or **1 vs rest** classifier

Multiclass classification: How to classify between K classes?

- **1 vs all** or **1 vs rest** classifier
- Use $K - 1$ classifiers, each solving a two class problem which separates a point in class k from point not in this class

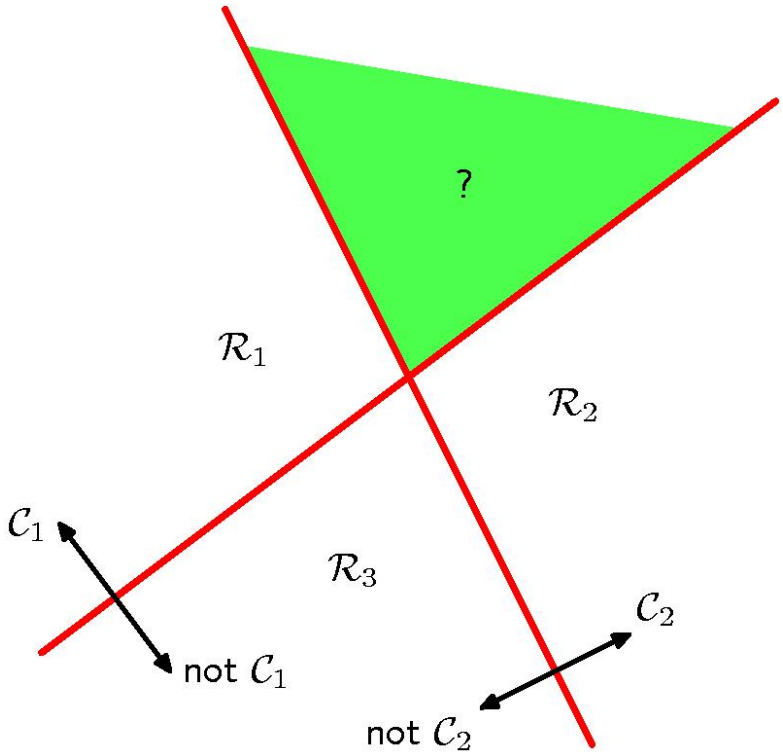
Multiclass classification: How to classify between K classes?

- **1 vs all** or **1 vs rest** classifier
- Use $K - 1$ classifiers, each solving a two class problem which separates a point in class k from point not in this class



Multiclass classification: How to classify between K classes?

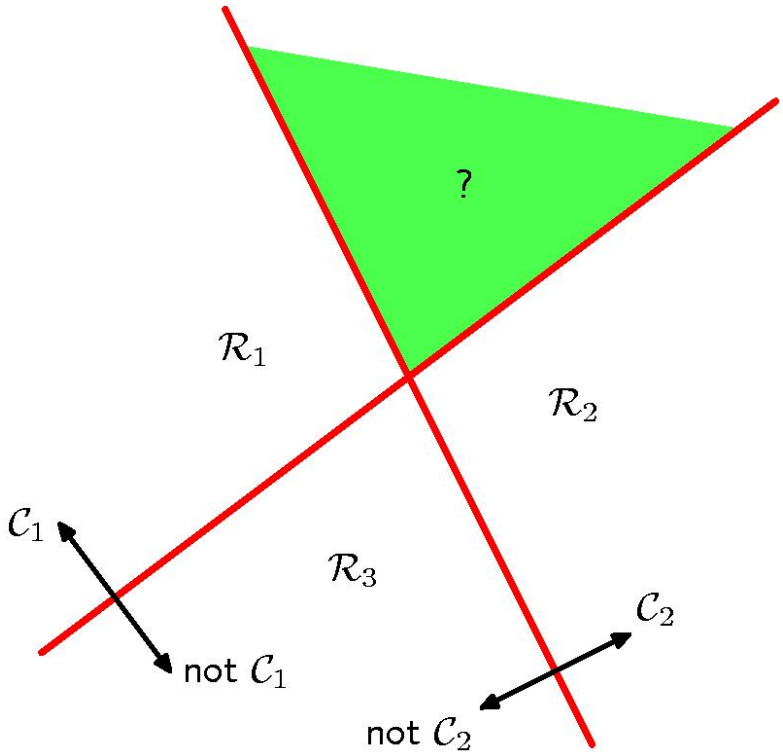
- **1 vs all** or **1 vs rest** classifier
- Use $K - 1$ classifiers, each solving a two class problem which separates a point in class k from point not in this class



- Issue:

Multiclass classification: How to classify between K classes?

- **1 vs all** or **1 vs rest** classifier
- Use $K - 1$ classifiers, each solving a two class problem which separates a point in class k from point not in this class



- Issue: more than one good answer or no good answer

Multiclass classification: How to classify between K classes?

Multiclass classification: How to classify between K classes?

- **1 vs 1** classifier

Multiclass classification: How to classify between K classes?

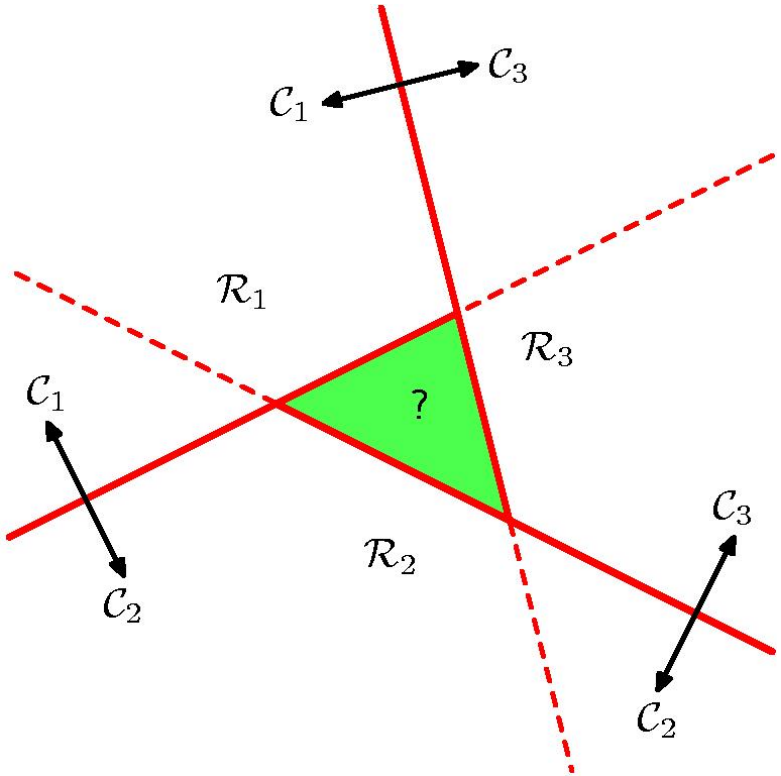
- **1 vs 1** classifier
- Use $K(K-1)/2$ two-way classifiers, one for each possible pair of classes

Multiclass classification: How to classify between K classes?

- **1 vs 1** classifier
- Use $K(K-1)/2$ two-way classifiers, one for each possible pair of classes
- Classify sample according to majority vote

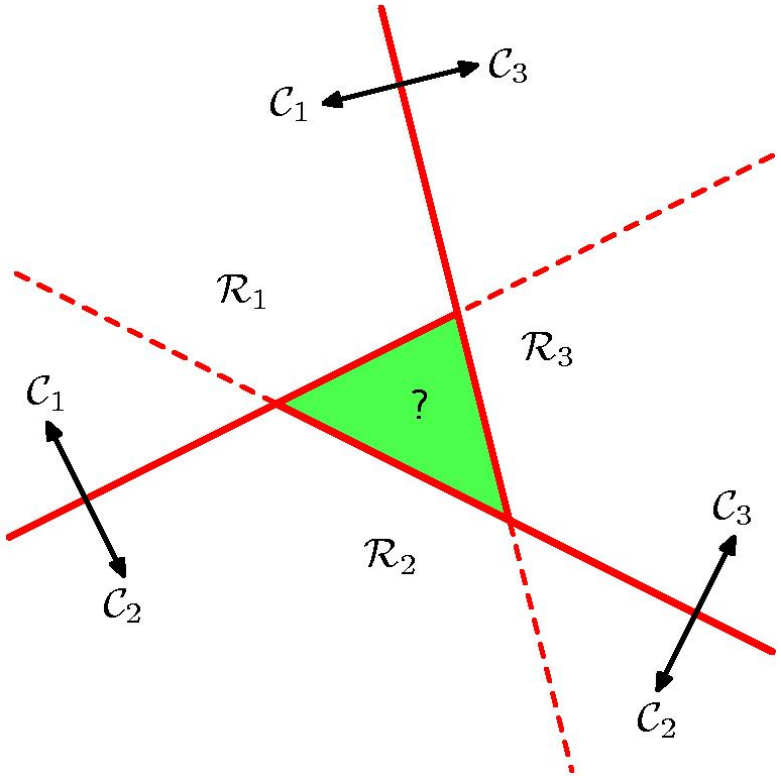
Multiclass classification: How to classify between K classes?

- **1 vs 1** classifier
- Use $K(K-1)/2$ two-way classifiers, one for each possible pair of classes
- Classify sample according to majority vote



Multiclass classification: How to classify between K classes?

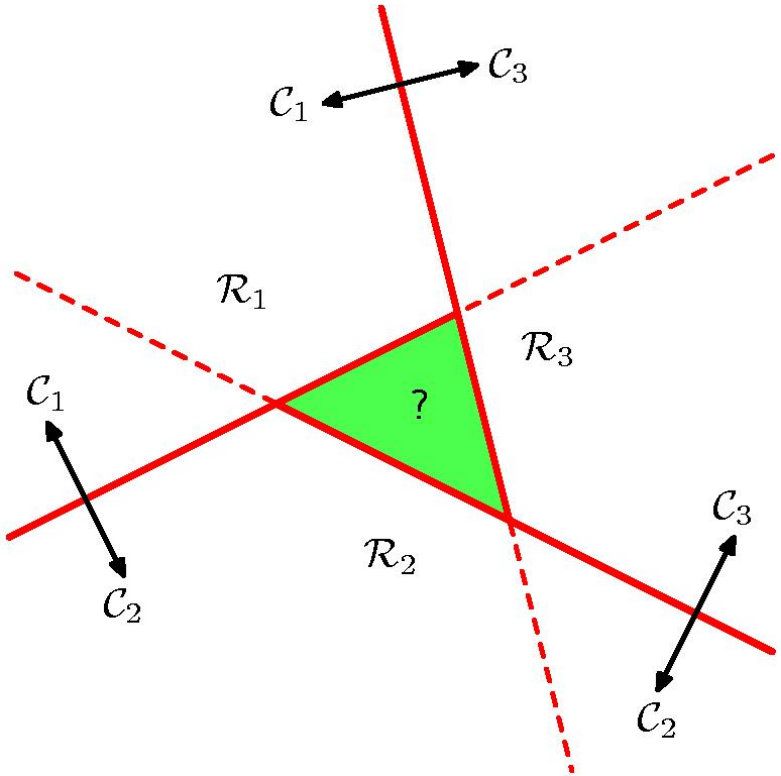
- **1 vs 1** classifier
- Use $K(K-1)/2$ two-way classifiers, one for each possible pair of classes
- Classify sample according to majority vote



- Issue:

Multiclass classification: How to classify between K classes?

- **1 vs 1** classifier
- Use $K(K-1)/2$ two-way classifiers, one for each possible pair of classes
- Classify sample according to majority vote



- Issue: two-way preferences need not be transitive

Multiclass classification: How to classify between K classes?

Multiclass classification: How to classify between K classes?

- Use a multinomial distribution over $y \in \{0, 1, \ldots, K-1\}$

Multiclass classification: How to classify between K classes?

- Use a multinomial distribution over $y \in \{0, 1, \ldots, K-1\}$
- Use K weight vectors $\boldsymbol{w}_{(y)}$

Multiclass classification: How to classify between K classes?

- Use a multinomial distribution over $y \in \{0, 1, \ldots, K-1\}$
- Use K weight vectors $\boldsymbol{w}_{(y)}$
- How to parameterize the multinomial distribution?

$$p(y = k|x^{(i)}) = \frac{\exp \boldsymbol{w}_{(k)}^\top \phi(x^{(i)})}{\sum_{j \in \{0,1,\ldots,K-1\}} \exp \boldsymbol{w}_{(j)}^\top \phi(x^{(i)})}$$

Multiclass classification: How to classify between K classes?

- Use a multinomial distribution over $y \in \{0, 1, \ldots, K-1\}$
- Use K weight vectors $\boldsymbol{w}_{(y)}$
- How to parameterize the multinomial distribution?

$$p(y = k|x^{(i)}) = \frac{\exp \boldsymbol{w}_{(k)}^\top \phi(x^{(i)})}{\sum_{j \in \{0,1,\ldots,K-1\}} \exp \boldsymbol{w}_{(j)}^\top \phi(x^{(i)})}$$

- We are using one parameter vector $\boldsymbol{w}_{(y)}$ per class

Multiclass classification: How to classify between K classes?

- Use a multinomial distribution over $y \in \{0, 1, \ldots, K-1\}$
- Use K weight vectors $\boldsymbol{w}_{(y)}$
- How to parameterize the multinomial distribution?

$$p(y = k|x^{(i)}) = \frac{\exp \boldsymbol{w}_{(k)}^\top \phi(x^{(i)})}{\sum_{j \in \{0,1,\ldots,K-1\}} \exp \boldsymbol{w}_{(j)}^\top \phi(x^{(i)})}$$

- We are using one parameter vector $\boldsymbol{w}_{(y)}$ per class
- Maximizing the likelihood as before

Multiclass classification: How to classify between K classes?

- Use a multinomial distribution over $y \in \{0, 1, \ldots, K-1\}$
- Use K weight vectors $\boldsymbol{w}_{(y)}$
- How to parameterize the multinomial distribution?

$$p(y = k|x^{(i)}) = \frac{\exp \boldsymbol{w}_{(k)}^\top \phi(x^{(i)})}{\sum_{j \in \{0,1,\ldots,K-1\}} \exp \boldsymbol{w}_{(j)}^\top \phi(x^{(i)})}$$

- We are using one parameter vector $\boldsymbol{w}_{(y)}$ per class
- Maximizing the likelihood as before

$$\arg\max_{\boldsymbol{w}} \prod_{(x^{(i)},y^{(i)}) \in \mathcal{D}} p(y = y^{(i)}|x^{(i)}) = \arg\min_{\boldsymbol{w}} \sum_{(x^{(i)},y^{(i)}) \in \mathcal{D}} -\log p(y = y^{(i)}|x^{(i)})$$

Questions:

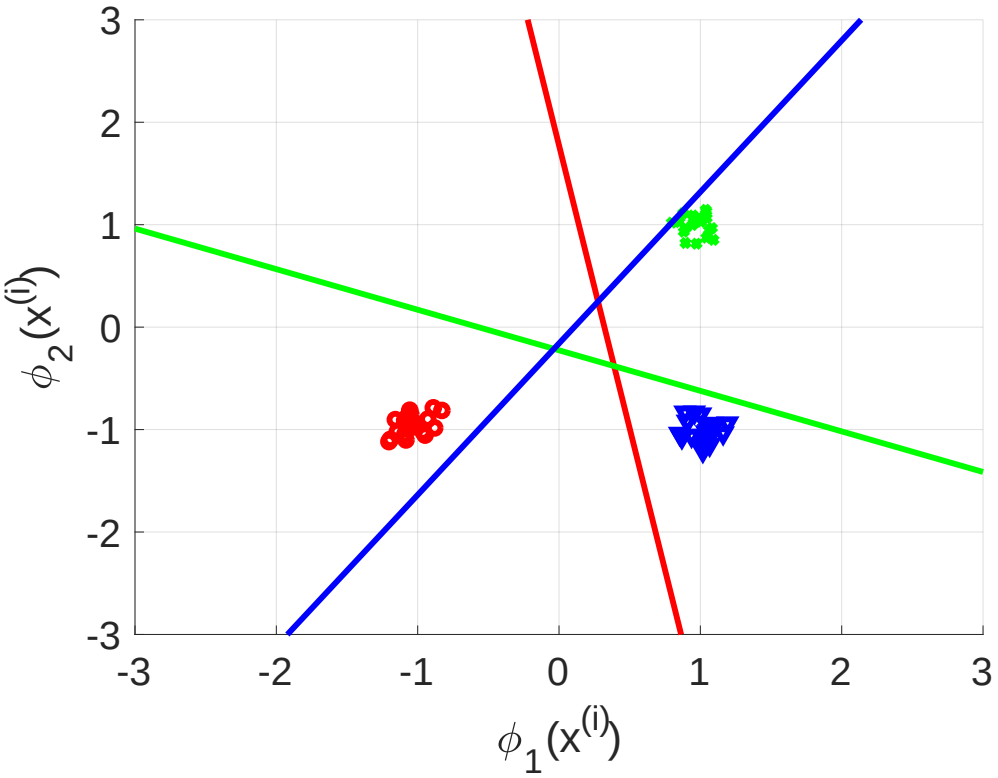
Questions:

- Example to build intuition?

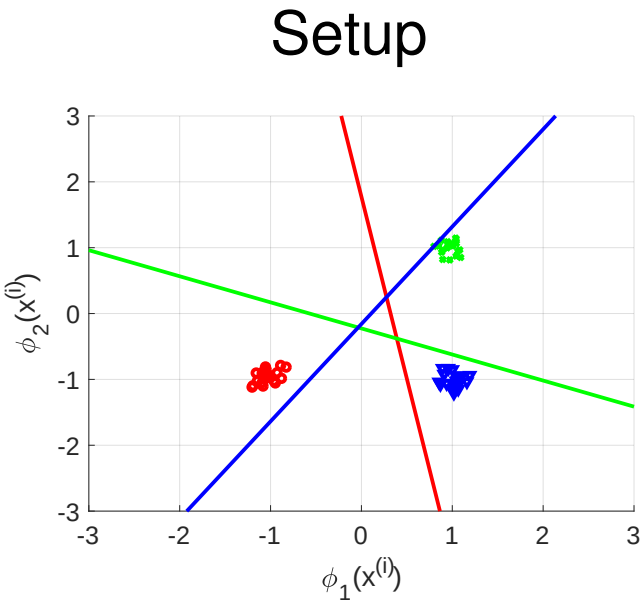
Questions:

- Example to build intuition?
- How does this relate to our earlier binary setting?

Example: $(\phi_3(x^{(i)}) = 1)$

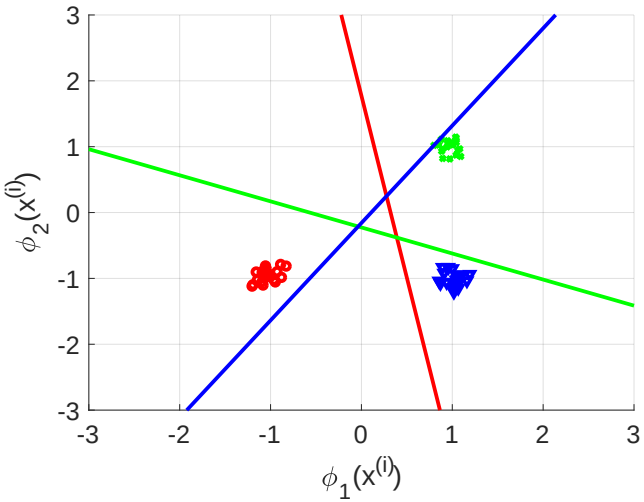


Example: $(\phi_3(x^{(i)}) = 1)$



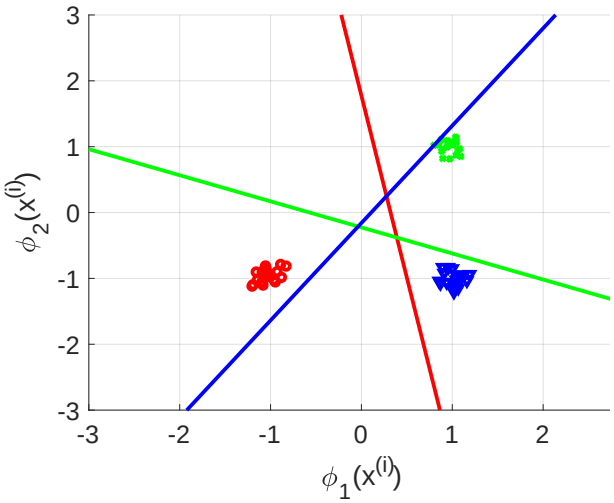
Example: $(\phi_3(x^{(i)}) = 1)$

Setup Decision Boundary

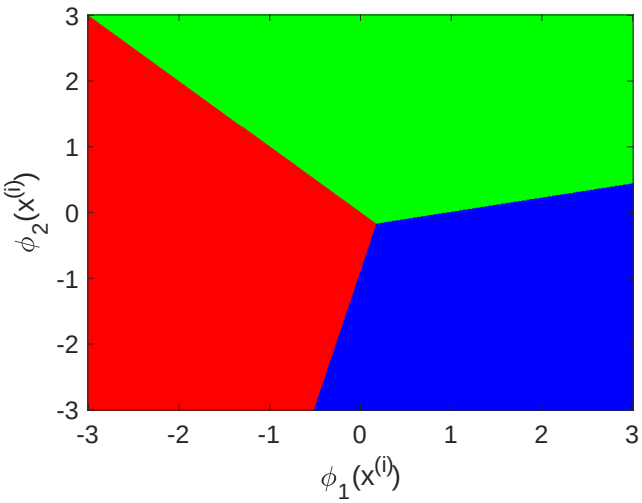


Example: $(\phi_3(x^{(i)}) = 1)$

Setup

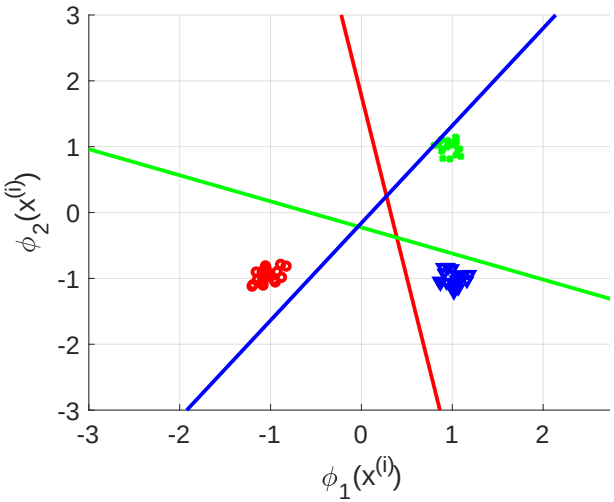


Decision Boundary

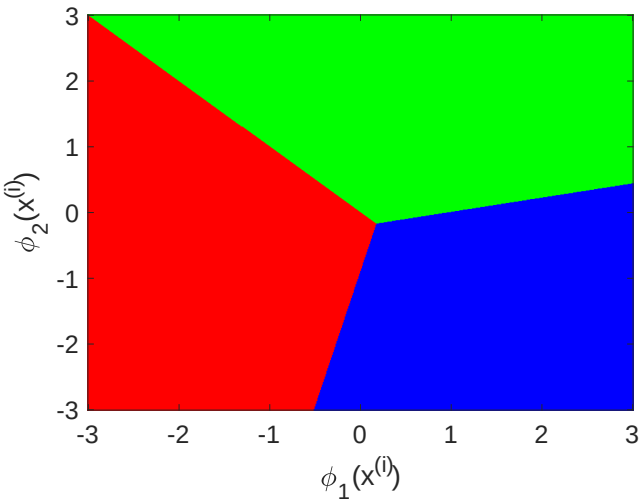


Example: $(\phi_3(x^{(i)}) = 1)$

Setup

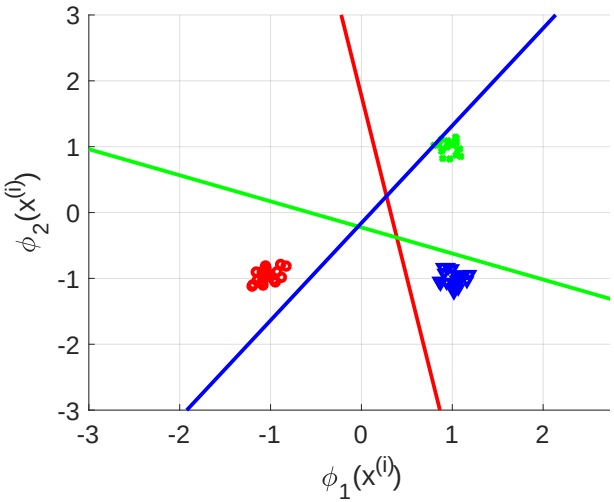


Decision Boundary Probability for $k = 1$

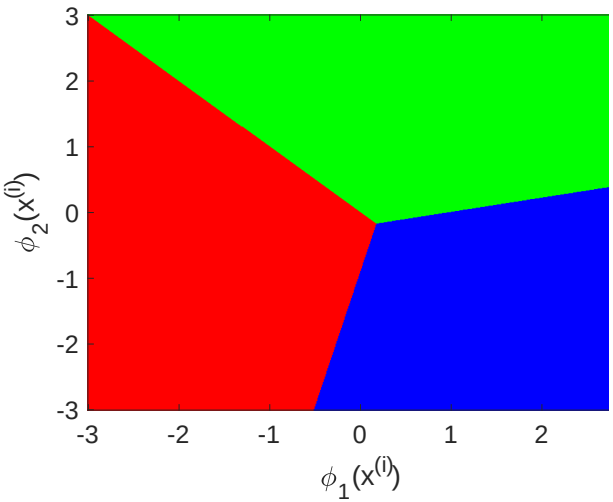


Example: $(\phi_3(x^{(i)}) = 1)$

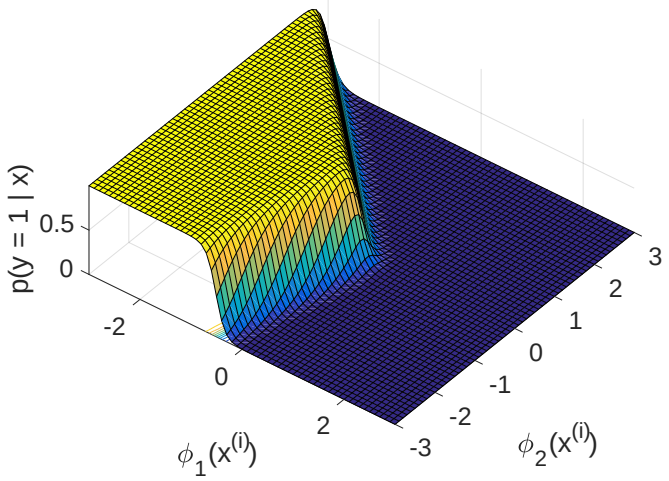
Setup



Decision Boundary



Probability for $k = 1$



Example: What happens when we have two classes?

Example: What happens when we have two classes?

- We get two decision boundaries. How do they related to binary logistic regression?

Example: What happens when we have two classes?

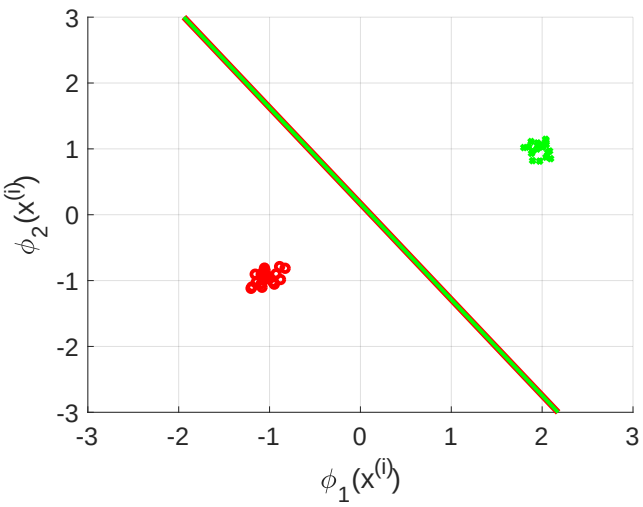
- We get two decision boundaries. How do they related to binary logistic regression?

Setup

Example: What happens when we have two classes?

- We get two decision boundaries. How do they related to binary logistic regression?

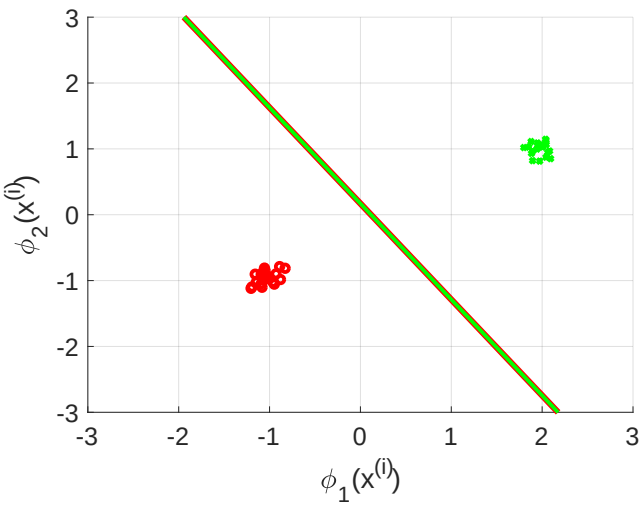
Setup



Example: What happens when we have two classes?

- We get two decision boundaries. How do they related to binary logistic regression?

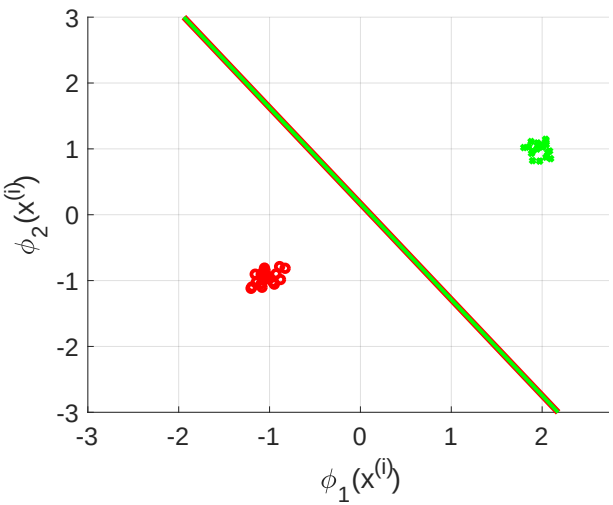
Setup Decision Boundary



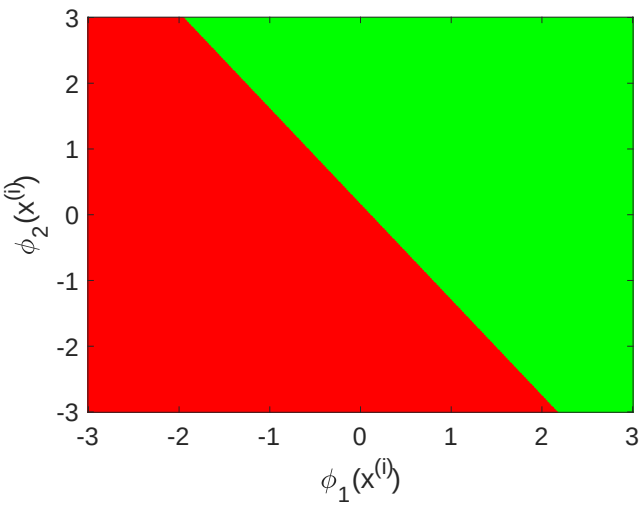
Example: What happens when we have two classes?

- We get two decision boundaries. How do they related to binary logistic regression?

Setup



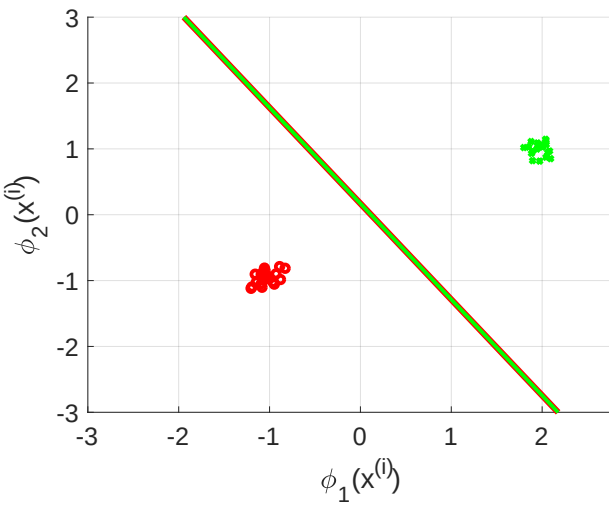
Decision Boundary



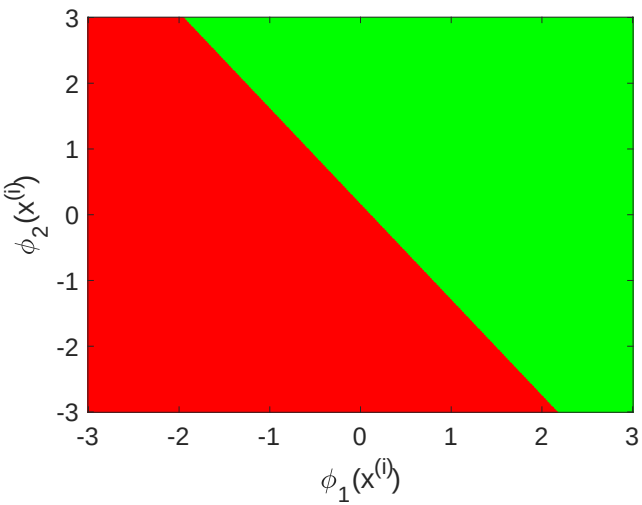
Example: What happens when we have two classes?

- We get two decision boundaries. How do they related to binary logistic regression?

Setup



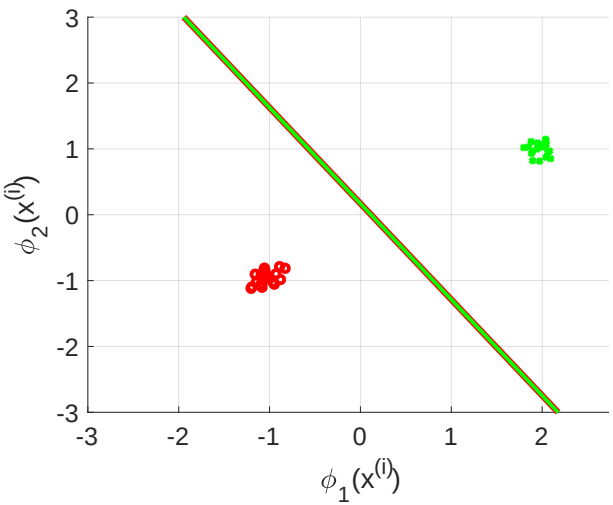
Decision Boundary Probability for $k = 1$



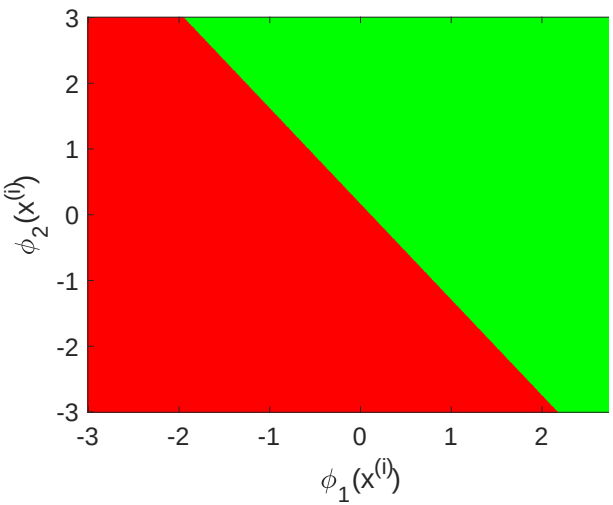
Example: What happens when we have two classes?

- We get two decision boundaries. How do they related to binary logistic regression?

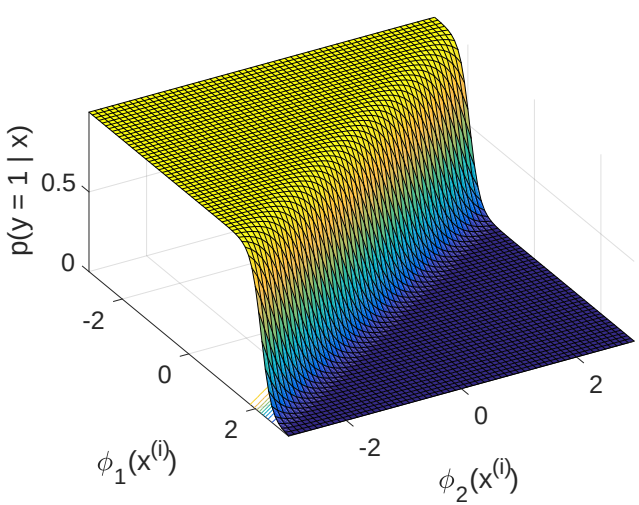
Setup



Decision Boundary



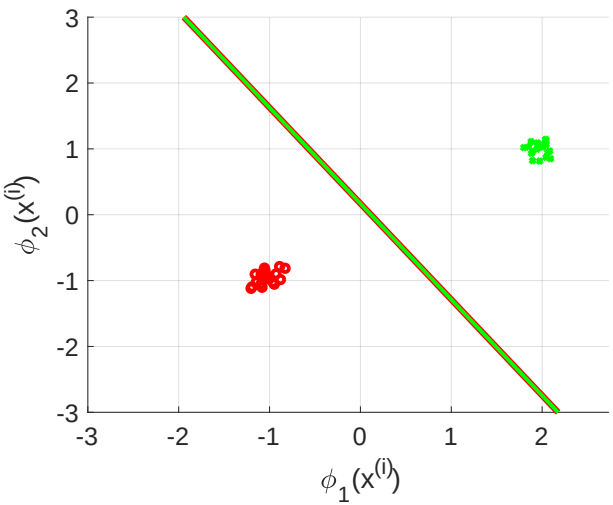
Probability for $k = 1$



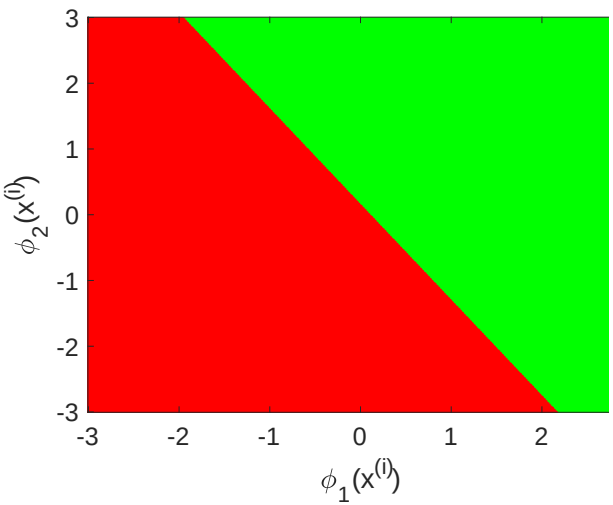
Example: What happens when we have two classes?

- We get two decision boundaries. How do they related to binary logistic regression?

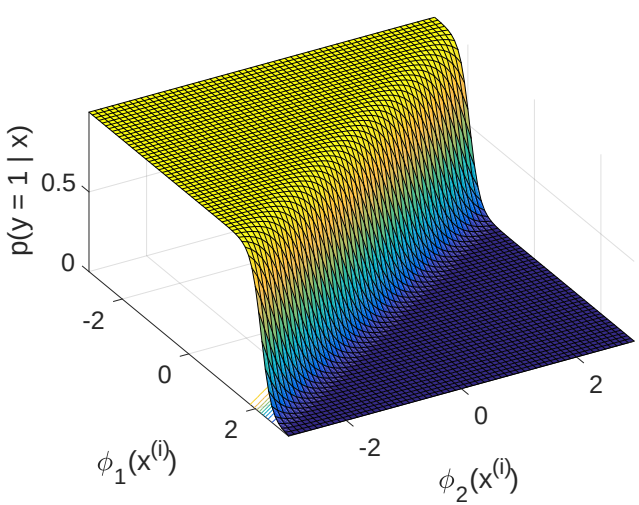
Setup



Decision Boundary



Probability for $k = 1$

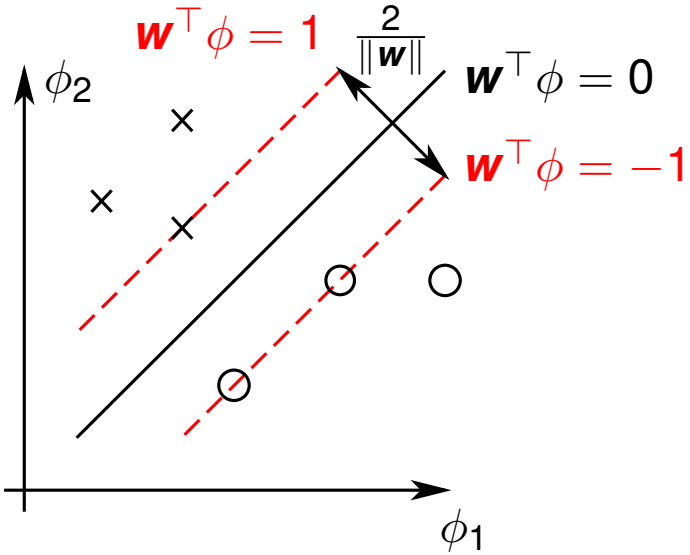


Lines coincide ($\boldsymbol{w}_{(1)} = -\boldsymbol{w}_{(2)}$)

How about multi-class SVM?

Recap: Binary SVM

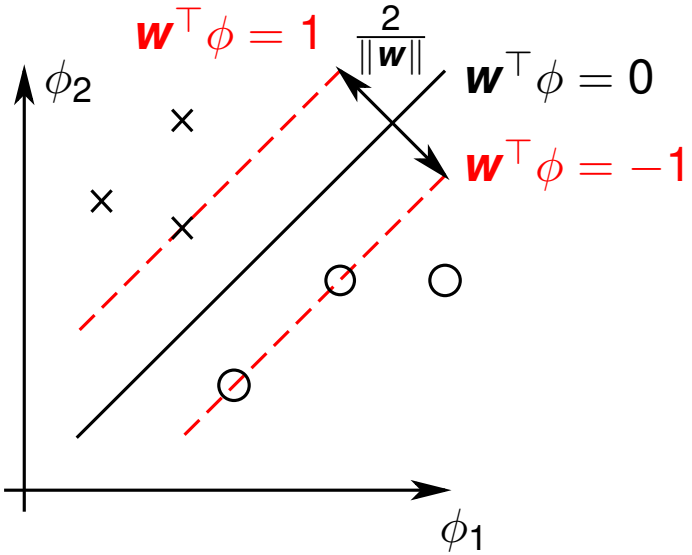
Intuitively:



How about multi-class SVM?

Recap: Binary SVM

Intuitively:



$$u_1 + c\boldsymbol{w} = u_2 \qquad \boldsymbol{w}^\top u_2 = -1$$

$$\boldsymbol{w}^\top (u_1 + c\boldsymbol{w}) = -1$$

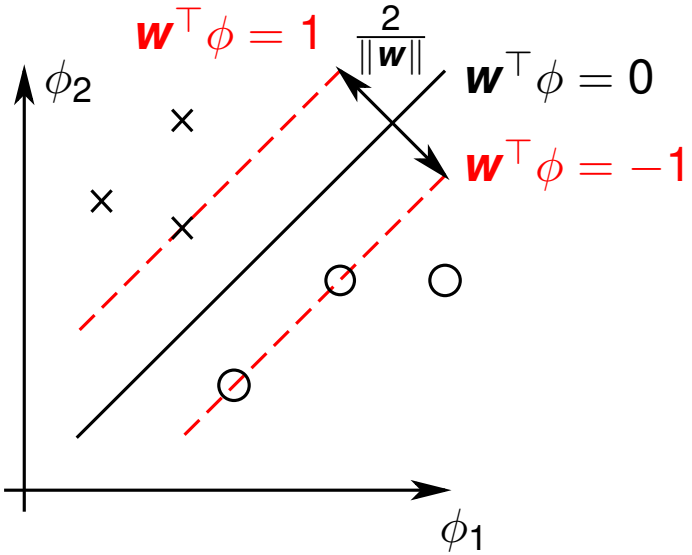
$$c = \frac{-2}{\|\boldsymbol{w}\|_2^2}$$

$$\|u_1 - u_2\|_2 = \| - c\boldsymbol{w}\|_2 = \frac{2}{\|\boldsymbol{w}\|_2^2}\|\boldsymbol{w}\| = \frac{2}{\|\boldsymbol{w}\|}$$

How about multi-class SVM?

Recap: Binary SVM

Intuitively:



$$u_1 + cw = u_2 \qquad w^\top u_2 = -1$$

$$w^\top (u_1 + cw) = -1$$

$$c = \frac{-2}{\|w\|_2^2}$$

$$\|u_1 - u_2\|_2 = \| - cw\|_2 = \frac{2}{\|w\|_2^2}\|w\| = \frac{2}{\|w\|}$$

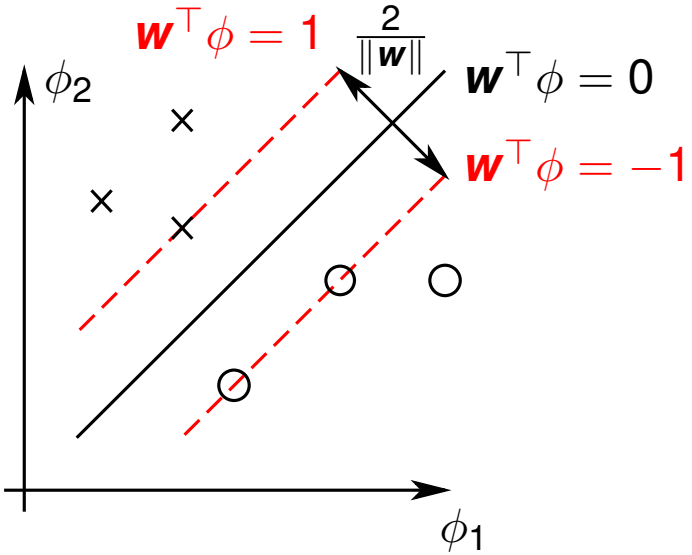
Maximize margin $\frac{2}{\|w\|}$:

$$\min_w \frac{1}{2}\|w\|_2^2 \qquad \text{s.t.} \qquad y^{(i)} w^\top \phi(x^{(i)}) \geq \underbrace{1}_{\text{Taskloss: } L} \qquad \forall (x^{(i)}, y^{(i)}) \in \mathcal{D}$$

How about multi-class SVM?

Recap: Binary SVM

Intuitively:



$$u_1 + cw = u_2 \qquad w^\top u_2 = -1$$

$$w^\top (u_1 + cw) = -1$$

$$c = \frac{-2}{\|w\|_2^2}$$

$$\|u_1 - u_2\|_2 = \| - cw\|_2 = \frac{2}{\|w\|_2^2}\|w\| = \frac{2}{\|w\|}$$

Maximize margin $\frac{2}{\|w\|}$:

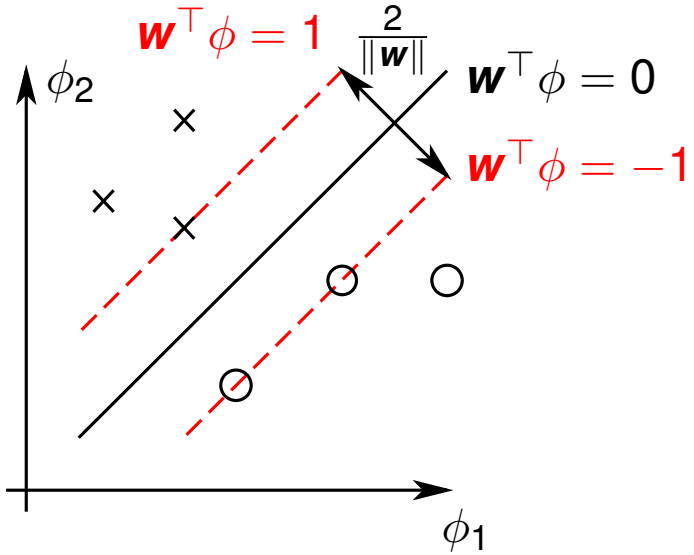
$$\min_w \frac{1}{2}\|w\|_2^2 \qquad \text{s.t.} \qquad y^{(i)} w^\top \phi(x^{(i)}) \geq \underbrace{1}_{\text{Taskloss: } L} \qquad \forall (x^{(i)}, y^{(i)}) \in \mathcal{D}$$

Note: any value $L \geq 0$ is okay.

How about multi-class SVM?

Recap: Binary SVM

Intuitively:



$$u_1 + cw = u_2 \qquad w^\top u_2 = -1$$

$$w^\top (u_1 + cw) = -1$$

$$c = \frac{-2}{\|w\|_2^2}$$

$$\|u_1 - u_2\|_2 = \| - cw\|_2 = \frac{2}{\|w\|_2^2}\|w\| = \frac{2}{\|w\|}$$

Maximize margin $\frac{2}{\|w\|}$:

$$\min_w \frac{1}{2}\|w\|_2^2 \qquad \text{s.t.} \qquad y^{(i)} w^\top \phi(x^{(i)}) \geq \underbrace{1}_{\text{Taskloss: } L} \qquad \forall (x^{(i)}, y^{(i)}) \in \mathcal{D}$$

Note: any value $L \geq 0$ is okay.

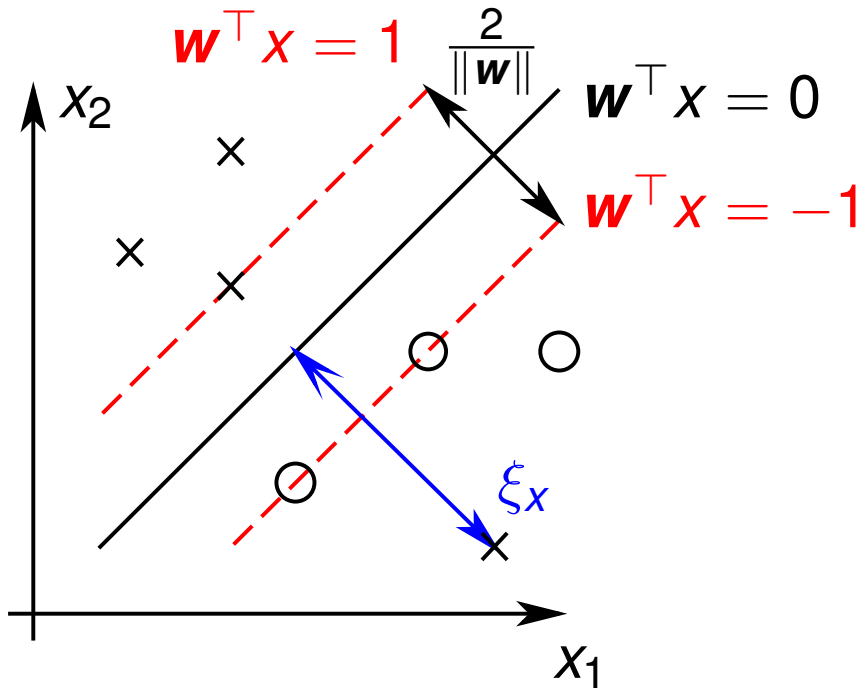
Issue: what if data not linearly separable?

Recap: Binary SVM

Introduce slack variables $\xi^{(i)}$:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \xi^{(i)} \qquad \text{s.t.} \qquad y^{(i)}\boldsymbol{w}^\top \phi(x^{(i)}) \geq 1 - \xi^{(i)} \quad \forall i \in \mathcal{D}$$

Intuitively:



How about multi-class SVM?

How about multi-class SVM? Recall $y \in \{0, 1, \ldots, K-1\}$ and class vectors $\boldsymbol{w}_{(y)}$

How about multi-class SVM? Recall $y \in \{0, 1, \ldots, K-1\}$ and class vectors $\boldsymbol{w}_{(y)}$

What do we want?

How about multi-class SVM? Recall $y \in \{0, 1, \ldots, K-1\}$ and class vectors $\boldsymbol{w}_{(y)}$

What do we want?

$$\boldsymbol{w}_{y^{(i)}}^T \phi(x^{(i)}) \geq \boldsymbol{w}_{\hat{y}}^T \phi(x^{(i)}) \qquad \forall i \in \mathcal{D}, \hat{y} \in \{0, 1, \ldots, K-1\}$$

How about multi-class SVM? Recall $y \in \{0, 1, \ldots, K-1\}$ and class vectors $\boldsymbol{w}_{(y)}$

What do we want?

$$\boldsymbol{w}_{y^{(i)}}^T \phi(x^{(i)}) \geq \boldsymbol{w}_{\hat{y}}^T \phi(x^{(i)}) \qquad \forall i \in \mathcal{D}, \hat{y} \in \{0, 1, \ldots, K-1\}$$

$$\min_{\boldsymbol{w}, \xi^{(i)} \geq 0} \frac{C}{2} \|\boldsymbol{w}\|_2^2 + \sum_{i \in \mathcal{D}} \xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}_{y^{(i)}}^T \phi(x^{(i)}) - \boldsymbol{w}_{\hat{y}}^T \phi(x^{(i)}) \geq 1 - \xi^{(i)} \quad \forall i, \hat{y}$$

Multiclass SVM objective:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}_{y^{(i)}}^T \phi(x^{(i)}) - \boldsymbol{w}_{\hat{y}}^T \phi(x^{(i)}) \geq 1 - \xi^{(i)} \quad \forall i, \hat{y}$$

Multiclass SVM objective:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}_{y^{(i)}}^T \phi(x^{(i)}) - \boldsymbol{w}_{\hat{y}}^T \phi(x^{(i)}) \geq 1 - \xi^{(i)} \quad \forall i, \hat{y}$$

Define mapping:

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_0 \\ \vdots \\ \boldsymbol{w}_{K-1} \end{bmatrix} \qquad \psi(x^{(i)}, y^{(i)}) = \begin{bmatrix} \phi(x^{(i)})\delta(y^{(i)} = 0) \\ \vdots \\ \phi(x^{(i)})\delta(y^{(i)} = K-1) \end{bmatrix}$$

Multiclass SVM objective:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}_{y^{(i)}}^T \phi(x^{(i)}) - \boldsymbol{w}_{\hat{y}}^T \phi(x^{(i)}) \geq 1 - \xi^{(i)} \quad \forall i, \hat{y}$$

Define mapping:

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_0 \\ \vdots \\ \boldsymbol{w}_{K-1} \end{bmatrix} \qquad \psi(x^{(i)}, y^{(i)}) = \begin{bmatrix} \phi(x^{(i)})\delta(y^{(i)} = 0) \\ \vdots \\ \phi(x^{(i)})\delta(y^{(i)} = K-1) \end{bmatrix}$$

Equivalent formulation:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}^T \left(\psi(x^{(i)}, y^{(i)}) - \psi(x^{(i)}, \hat{y})\right) \geq 1 - \xi^{(i)} \quad \forall i, \hat{y}$$

Multiclass SVM objective:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}^T\left(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})\right)\geq 1-\xi^{(i)} \quad \forall i,\hat{y}$$

Let's get rid of the slack variable:

Multiclass SVM objective:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}^T\left(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})\right)\geq 1-\xi^{(i)} \quad \forall i,\hat{y}$$

Let's get rid of the slack variable:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\max\{0,\max_{\hat{y}}(1-\boldsymbol{w}^T(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})))\}$$

Multiclass SVM objective:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}^T\left(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})\right)\geq 1-\xi^{(i)} \quad \forall i,\hat{y}$$

Let's get rid of the slack variable:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\max\{0,\underbrace{\max_{\hat{y}}(1-\boldsymbol{w}^T(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})))}_{\geq 0}\}$$

Multiclass SVM objective:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}^T\left(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})\right)\geq 1-\xi^{(i)} \quad \forall i,\hat{y}$$

Let's get rid of the slack variable:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\max\{0,\underbrace{\max_{\hat{y}}(1-\boldsymbol{w}^T(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})))}_{\geq 0}\}$$

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\max_{\hat{y}}(1-\boldsymbol{w}^T(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})))$$

Multiclass SVM objective:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}^T\left(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})\right)\geq 1-\xi^{(i)} \quad \forall i,\hat{y}$$

Let's get rid of the slack variable:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\max\{0,\underbrace{\max_{\hat{y}}(1-\boldsymbol{w}^T(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})))}_{\geq 0}\}$$

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\max_{\hat{y}}(1-\boldsymbol{w}^T(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})))$$

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\max_{\hat{y}}(1+\boldsymbol{w}^T\psi(x^{(i)},\hat{y}))-\boldsymbol{w}^T\psi(x^{(i)},y^{(i)})$$

Multiclass SVM objective:

$$\min_{\boldsymbol{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\xi^{(i)} \quad \text{s.t.} \quad \boldsymbol{w}^T\left(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})\right)\geq 1-\xi^{(i)} \quad \forall i,\hat{y}$$

Let's get rid of the slack variable:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\max\{0,\underbrace{\max_{\hat{y}}(1-\boldsymbol{w}^T(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})))}_{\geq 0}\}$$

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\max_{\hat{y}}(1-\boldsymbol{w}^T(\psi(x^{(i)},y^{(i)})-\psi(x^{(i)},\hat{y})))$$

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2+\sum_{i\in\mathcal{D}}\underbrace{\max_{\hat{y}}(1+\boldsymbol{w}^T\psi(x^{(i)},\hat{y}))}_{\text{Loss-augmented inference}}-\boldsymbol{w}^T\psi(x^{(i)},y^{(i)})$$

How does multiclass SVM relate to multiclass logistic regression?

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \max_{\hat{y}}(1 + \boldsymbol{w}^T\psi(x^{(i)}, \hat{y})) - \boldsymbol{w}^T\psi(x^{(i)}, y^{(i)})$$

How does multiclass SVM relate to multiclass logistic regression?

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \underbrace{\max_{\hat{y}}(1 + \boldsymbol{w}^T\psi(x^{(i)}, \hat{y}))}_{\text{Loss-augmented inference}} - \boldsymbol{w}^T\psi(x^{(i)}, y^{(i)})$$

How does multiclass SVM relate to multiclass logistic regression?

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \underbrace{\max_{\hat{y}}(1 + \boldsymbol{w}^T\psi(x^{(i)}, \hat{y}))}_{\text{Loss-augmented inference}} - \boldsymbol{w}^T\psi(x^{(i)}, y^{(i)})$$

Same temperature argument as in the last lecture.

How does multiclass SVM relate to multiclass logistic regression?

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \underbrace{\max_{\hat{y}}(1 + \boldsymbol{w}^T\psi(x^{(i)}, \hat{y}))}_{\text{Loss-augmented inference}} - \boldsymbol{w}^T\psi(x^{(i)}, y^{(i)})$$

Same temperature argument as in the last lecture.

Homework

How does multiclass SVM relate to multiclass logistic regression?

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \underbrace{\max_{\hat{y}}(1 + \boldsymbol{w}^T\psi(x^{(i)},\hat{y}))}_{\text{Loss-augmented inference}} - \boldsymbol{w}^T\psi(x^{(i)},y^{(i)})$$

Same temperature argument as in the last lecture.

Homework

Solution:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \epsilon \ln \sum_{\hat{y}} \exp \frac{L(y^{(i)},\hat{y}) + \boldsymbol{w}^T\psi(x^{(i)},\hat{y}))}{\epsilon} - \boldsymbol{w}^T\psi(x^{(i)},y^{(i)})$$

Our current framework:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \epsilon \ln \sum_{\hat{y}} \exp \frac{L(y^{(i)},\hat{y}) + \boldsymbol{w}^T\psi(x^{(i)},\hat{y}))}{\epsilon} - \boldsymbol{w}^T\psi(x^{(i)},y^{(i)})$$

Our current framework:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \epsilon \ln \sum_{\hat{y}} \exp \frac{L(y^{(i)},\hat{y}) + \boldsymbol{w}^T\psi(x^{(i)},\hat{y}))}{\epsilon} - \boldsymbol{w}^T\psi(x^{(i)},y^{(i)})$$

What is a possible issue/limitation?

Our current framework:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i \in \mathcal{D}} \epsilon \ln \sum_{\hat{y}} \exp \frac{L(y^{(i)}, \hat{y}) + \boldsymbol{w}^T \psi(x^{(i)}, \hat{y}))}{\epsilon} - \boldsymbol{w}^T \psi(x^{(i)}, y^{(i)})$$

What is a possible issue/limitation?

Linearity in the feature space $\psi(x^{(i)}, y^{(i)})$.

Our current framework:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \epsilon \ln \sum_{\hat{y}} \exp \frac{L(y^{(i)},\hat{y}) + \boldsymbol{w}^T\psi(x^{(i)},\hat{y}))}{\epsilon} - \boldsymbol{w}^T\psi(x^{(i)},y^{(i)})$$

What is a possible issue/limitation?

Linearity in the feature space $\psi(x^{(i)}, y^{(i)})$.

How to fix this?

Our current framework:

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i\in\mathcal{D}} \epsilon \ln \sum_{\hat{y}} \exp \frac{L(y^{(i)},\hat{y}) + \boldsymbol{w}^T\psi(x^{(i)},\hat{y}))}{\epsilon} - \boldsymbol{w}^T\psi(x^{(i)},y^{(i)})$$

What is a possible issue/limitation?

Linearity in the feature space $\psi(x^{(i)}, y^{(i)})$.

How to fix this?

Use Kernels

Dual of Logistic Regression:

Dual of Logistic Regression:

$$\max_{0 \leq \lambda^{(i)} \leq 1} g(\lambda) := \frac{-1}{2C} \| \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i H(\lambda^{(i)})$$

Dual of Logistic Regression:

$$\max_{0 \leq \lambda^{(i)} \leq 1} g(\lambda) := \frac{-1}{2C} \| \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i H(\lambda^{(i)})$$

Prediction with dual variables:

Dual of Logistic Regression:

$$\max_{0 \leq \lambda^{(i)} \leq 1} g(\lambda) := \frac{-1}{2C} \| \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i H(\lambda^{(i)})$$

Prediction with dual variables:

$$\boldsymbol{w}^\top \phi(x) = \frac{1}{C} \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)})^\top \phi(x)$$

Dual of Logistic Regression:

$$\max_{0 \leq \lambda^{(i)} \leq 1} g(\lambda) := \frac{-1}{2C} \| \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i H(\lambda^{(i)})$$

Prediction with dual variables:

$$\boldsymbol{w}^\top \phi(x) = \frac{1}{C} \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)})^\top \phi(x)$$

Dual of SVM:

Dual of Logistic Regression:

$$\max_{0 \leq \lambda^{(i)} \leq 1} g(\lambda) := \frac{-1}{2C} \| \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i H(\lambda^{(i)})$$

Prediction with dual variables:

$$\boldsymbol{w}^\top \phi(x) = \frac{1}{C} \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)})^\top \phi(x)$$

Dual of SVM:

$$\max_{0 \leq \alpha \leq 1} g(\alpha) := \frac{-1}{2C} \| \sum_i \alpha^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i \alpha^{(i)}$$

Dual of Logistic Regression:

$$\max_{0 \leq \lambda^{(i)} \leq 1} g(\lambda) := \frac{-1}{2C} \| \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i H(\lambda^{(i)})$$

Prediction with dual variables:

$$\boldsymbol{w}^\top \phi(x) = \frac{1}{C} \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)})^\top \phi(x)$$

Dual of SVM:

$$\max_{0 \leq \alpha \leq 1} g(\alpha) := \frac{-1}{2C} \| \sum_i \alpha^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i \alpha^{(i)}$$

Prediction with dual variables:

Dual of Logistic Regression:

$$\max_{0 \leq \lambda^{(i)} \leq 1} g(\lambda) := \frac{-1}{2C} \| \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i H(\lambda^{(i)})$$

Prediction with dual variables:

$$\boldsymbol{w}^\top \phi(x) = \frac{1}{C} \sum_i \lambda^{(i)} y^{(i)} \phi(x^{(i)})^\top \phi(x)$$

Dual of SVM:

$$\max_{0 \leq \alpha \leq 1} g(\alpha) := \frac{-1}{2C} \| \sum_i \alpha^{(i)} y^{(i)} \phi(x^{(i)}) \|_2^2 + \sum_i \alpha^{(i)}$$

Prediction with dual variables:

$$\boldsymbol{w}^\top \phi(x) = \frac{1}{C} \sum_i \alpha^{(i)} y^{(i)} \phi(x^{(i)})^\top \phi(x)$$

Observation:

Observation:

Many inner products $\phi(x^{(i)})^\top \phi(x^{(j)})$ between different samples

Observation:

Many inner products $\phi(x^{(i)})^\top \phi(x^{(j)})$ between different samples

Kernel trick:

Observation:

Many inner products $\phi(x^{(i)})^\top \phi(x^{(j)})$ between different samples

Kernel trick:

Replace inner products with call to kernel $\kappa(x^{(i)}, x^{(j)}) \in \mathbb{R}$

Observation:

Many inner products $\phi(x^{(i)})^\top \phi(x^{(j)})$ between different samples

Kernel trick:

Replace inner products with call to kernel $\kappa(x^{(i)}, x^{(j)}) \in \mathbb{R}$

Advantage:

Observation:

Many inner products $\phi(x^{(i)})^\top \phi(x^{(j)})$ between different samples

Kernel trick:

Replace inner products with call to kernel $\kappa(x^{(i)}, x^{(j)}) \in \mathbb{R}$

Advantage:

No need to explicitly construct feature vector $\phi(x)$

Observation:

Many inner products $\phi(x^{(i)})^\top \phi(x^{(j)})$ between different samples

Kernel trick:

Replace inner products with call to kernel $\kappa(x^{(i)}, x^{(j)}) \in \mathbb{R}$

Advantage:

No need to explicitly construct feature vector $\phi(x)$

But:

Observation:

Many inner products $\phi(x^{(i)})^\top \phi(x^{(j)})$ between different samples

Kernel trick:

Replace inner products with call to kernel $\kappa(x^{(i)}, x^{(j)}) \in \mathbb{R}$

Advantage:

No need to explicitly construct feature vector $\phi(x)$

But:

We need to ensure that κ is a valid kernel, i.e., it corresponds to an inner product in some feature space

Example:

$$\kappa(x, z) \quad = \quad (x^\top z)^2$$

Example:

$$\begin{aligned}\kappa(x,z) &= (x^\top z)^2 \\ &= x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + 2x_2^2 z_2^2\end{aligned}$$

Example:

$$
\begin{aligned}
\kappa(x,z) &= (x^\top z)^2 \\
&= x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + 2 x_2^2 z_2^2 \\
&= \begin{bmatrix} x_1^2 & \sqrt{2} x_1 x_2 & x_2^2 \end{bmatrix} \begin{bmatrix} z_1^2 & \sqrt{2} z_1 z_2 & z_2^2 \end{bmatrix}^\top
\end{aligned}
$$

Example:

$$
\begin{aligned}
\kappa(x,z) &= (x^\top z)^2 \\
&= x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + 2x_2^2 z_2^2 \\
&= [\ x_1^2 \quad \sqrt{2}x_1 x_2 \quad x_2^2\ ][\ z_1^2 \quad \sqrt{2}z_1 z_2 \quad z_2^2\ ]^\top \\
&= \phi(x)^\top \phi(z)
\end{aligned}
$$

Example:

$$\begin{aligned}
\kappa(x,z) &= (x^\top z)^2 \\
&= x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + 2x_2^2 z_2^2 \\
&= [\ x_1^2 \quad \sqrt{2}x_1 x_2 \quad x_2^2\ ][\ z_1^2 \quad \sqrt{2}z_1 z_2 \quad z_2^2\ ]^\top \\
&= \phi(x)^\top \phi(z)
\end{aligned}$$

How to test without explicitly constructing feature vectors:

Example:

$$\begin{aligned}
\kappa(x,z) &= (x^\top z)^2 \\
&= x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + 2x_2^2 z_2^2 \\
&= [\ x_1^2 \quad \sqrt{2}x_1 x_2 \quad x_2^2\ ][\ z_1^2 \quad \sqrt{2}z_1 z_2 \quad z_2^2\ ]^\top \\
&= \phi(x)^\top \phi(z)
\end{aligned}$$

How to test without explicitly constructing feature vectors:

- Gram matrix $(\boldsymbol{K})_{i,j} = \kappa(x^{(i)}, x^{(j)})$ is positive definite $\forall x^{(i)}, x^{(j)}$

Example:

$$\begin{aligned}
\kappa(x,z) &= (x^\top z)^2 \\
&= x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + 2x_2^2 z_2^2 \\
&= [\ x_1^2 \quad \sqrt{2}x_1 x_2 \quad x_2^2\ ][\ z_1^2 \quad \sqrt{2}z_1 z_2 \quad z_2^2\ ]^\top \\
&= \phi(x)^\top \phi(z)
\end{aligned}$$

How to test without explicitly constructing feature vectors:

- Gram matrix $(\boldsymbol{K})_{i,j} = \kappa(x^{(i)}, x^{(j)})$ is positive definite $\forall x^{(i)}, x^{(j)}$
- Decompose/Compose kernel into/from known kernels

Example kernels:

Example kernels:

- Linear kernel:

$$\kappa(x^{(i)}, x^{(j)}) = {x^{(i)}}^\top x^{(j)}$$

Example kernels:

- Linear kernel:

$$\kappa(x^{(i)}, x^{(j)}) = {x^{(i)}}^\top x^{(j)}$$

- Squared exponential (Gaussian) kernel:

$$\kappa(x^{(i)}, x^{(j)}) = \exp\left(-\frac{1}{2}(x^{(i)} - x^{(j)})^\top \Sigma^{-1}(x^{(i)} - x^{(j)})\right)$$

Example kernels:

- Linear kernel:

$$\kappa(x^{(i)}, x^{(j)}) = {x^{(i)}}^\top x^{(j)}$$

- Squared exponential (Gaussian) kernel:

$$\kappa(x^{(i)}, x^{(j)}) = \exp\left(-\frac{1}{2}(x^{(i)} - x^{(j)})^\top \Sigma^{-1}(x^{(i)} - x^{(j)})\right)$$

- Sigmoid kernel:

$$\kappa(x^{(i)}, x^{(j)}) = \tanh(a \cdot {x^{(i)}}^\top x^{(j)} + b)$$

Techniques for constructing new kernels ($c > 0$):

Techniques for constructing new kernels ($c > 0$):

- Positive scaling:

$$\kappa(x^{(i)}, x^{(j)}) = c\kappa_1(x^{(i)}, x^{(j)})$$

Techniques for constructing new kernels ($c > 0$):

- Positive scaling:

$$\kappa(x^{(i)}, x^{(j)}) = c\kappa_1(x^{(i)}, x^{(j)})$$

- Exponentiation:

$$\kappa(x^{(i)}, x^{(j)}) = \exp(\kappa_1(x^{(i)}, x^{(j)}))$$

Techniques for constructing new kernels ($c > 0$):

- Positive scaling:

$$\kappa(x^{(i)}, x^{(j)}) = c\kappa_1(x^{(i)}, x^{(j)})$$

- Exponentiation:

$$\kappa(x^{(i)}, x^{(j)}) = \exp(\kappa_1(x^{(i)}, x^{(j)}))$$

- Addition:

$$\kappa(x^{(i)}, x^{(j)}) = \kappa_1(x^{(i)}, x^{(j)}) + \kappa_2(x^{(i)}, x^{(j)})$$

Techniques for constructing new kernels ($c > 0$):

- Positive scaling:

$$\kappa(x^{(i)}, x^{(j)}) = c\kappa_1(x^{(i)}, x^{(j)})$$

- Exponentiation:

$$\kappa(x^{(i)}, x^{(j)}) = \exp(\kappa_1(x^{(i)}, x^{(j)}))$$

- Addition:

$$\kappa(x^{(i)}, x^{(j)}) = \kappa_1(x^{(i)}, x^{(j)}) + \kappa_2(x^{(i)}, x^{(j)})$$

- Multiplication with function:

$$\kappa(x^{(i)}, x^{(j)}) = f(x^{(i)})\kappa_1(x^{(i)}, x^{(j)})f(x^{(j)})$$

Techniques for constructing new kernels ($c > 0$):

- Positive scaling:

$$\kappa(x^{(i)}, x^{(j)}) = c\kappa_1(x^{(i)}, x^{(j)})$$

- Exponentiation:

$$\kappa(x^{(i)}, x^{(j)}) = \exp(\kappa_1(x^{(i)}, x^{(j)}))$$

- Addition:

$$\kappa(x^{(i)}, x^{(j)}) = \kappa_1(x^{(i)}, x^{(j)}) + \kappa_2(x^{(i)}, x^{(j)})$$

- Multiplication with function:

$$\kappa(x^{(i)}, x^{(j)}) = f(x^{(i)})\kappa_1(x^{(i)}, x^{(j)})f(x^{(j)})$$

- Multiplication:

$$\kappa(x^{(i)}, x^{(j)}) = \kappa_1(x^{(i)}, x^{(j)})\kappa_2(x^{(i)}, x^{(j)})$$

Quiz:

Quiz:

- How to extend binary classification to multiple classes?

Quiz:

- How to extend binary classification to multiple classes?
- What are kernels good for?

Important topics of this lecture

- Multi-class classification
- Kernels as non-linear feature mappings

Up next:

- Deep nets