

CS 446/ECE 449: Machine Learning

A. G. Schwing

University of Illinois at Urbana-Champaign, 2020

Scribe & Exercises

L7: Multiclass Classification and Kernel Methods

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:

- Linear regression:

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

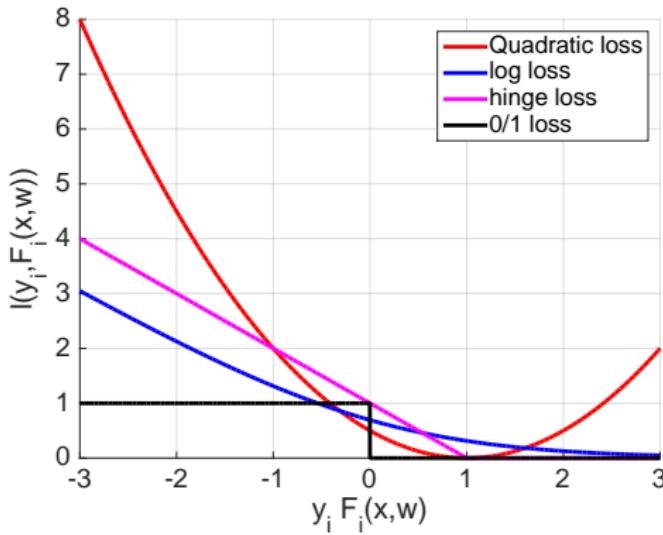
- Logistic regression:

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

- Binary SVM:

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

Recap:



Other loss functions:

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

Combining log- and hinge-loss:

$$\lim_{\epsilon \rightarrow 0} \epsilon \log \left(1 + \exp \frac{-F}{\epsilon} \right) =$$
$$\begin{aligned} & \stackrel{F \geq 0}{\equiv} 0 \\ & \stackrel{F \leq 0}{\equiv} \lim_{\epsilon \rightarrow 0} \frac{\log \left(1 + \exp \frac{-F}{\epsilon} \right)}{1/\epsilon} \quad \text{洛必达法则} \\ & = \lim_{\epsilon \rightarrow 0} \frac{\frac{\exp \frac{-F}{\epsilon}}{1 + \exp \frac{-F}{\epsilon}} \cdot (F/\epsilon^2)}{-1/\epsilon^2} \\ & = \lim_{\epsilon \rightarrow 0} \frac{1}{1 + \exp \frac{F}{\epsilon}} \cdot (-F) \\ & = -F \end{aligned}$$

In summary:

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

SVM as 0-temperature limit of logistic regression

Recap:

- Linear regression:

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

- Logistic regression:

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

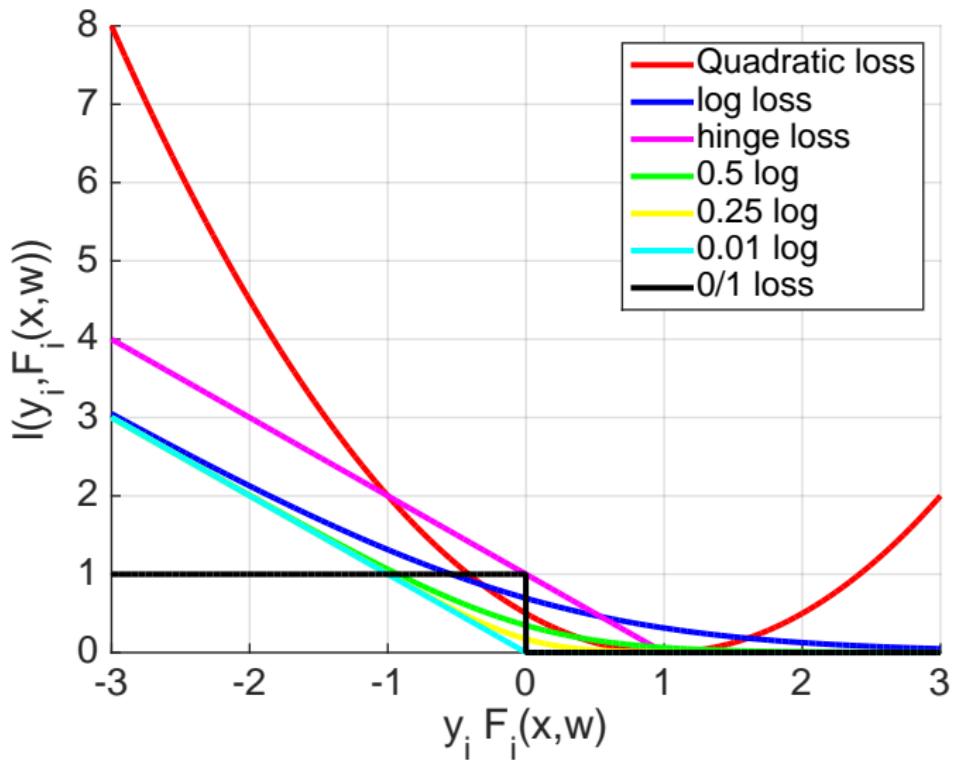
- Binary SVM:

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

- General binary classification:

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

Different loss functions



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?



Which object is illustrated?

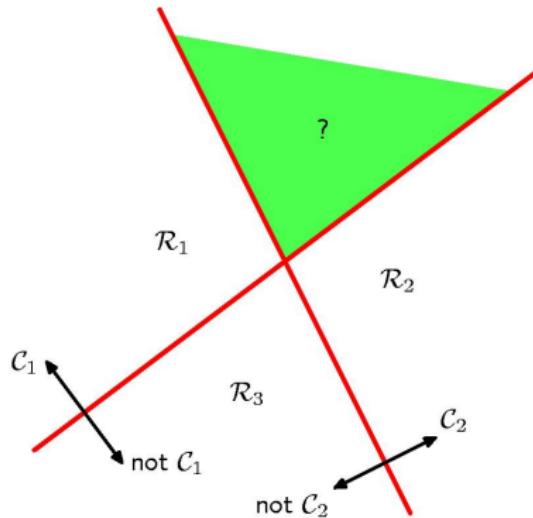
- Car
- Truck
- Recreational Vehicle
- Ambulance truck
- Fire truck

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?

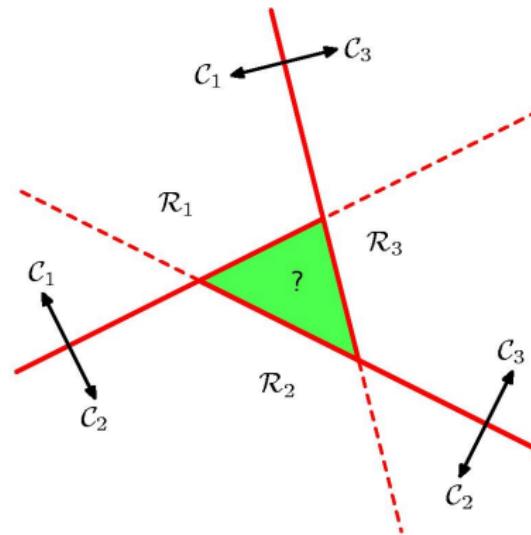
- **1 vs all** or **1 vs rest** classifier
- Use $K - 1$ classifiers, each solving a two class problem which separates a point in class k from point not in this class



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

Multiclass classification: How to classify between K classes?

- **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?

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

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

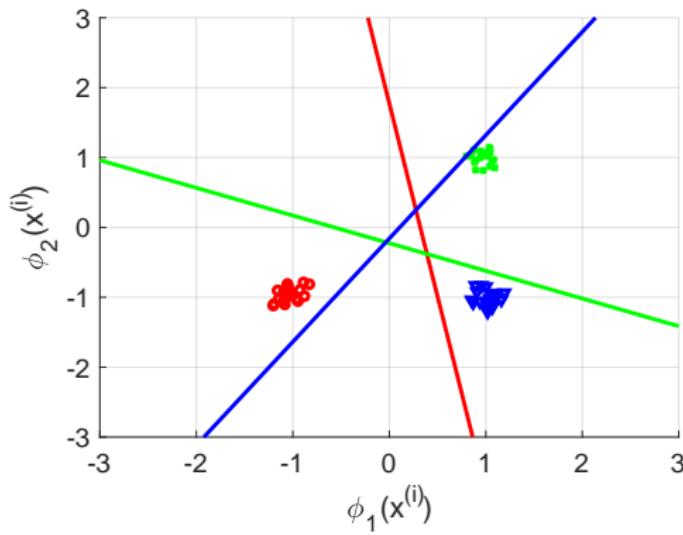
- We are using one parameter vector $\mathbf{w}_{(y)}$ per class
- Maximizing the likelihood as before

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

Questions:

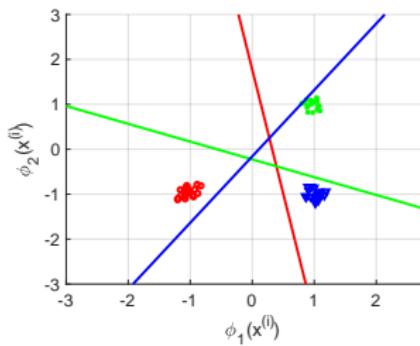
- Example to build intuition?
- How does this relate to our earlier binary setting?

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

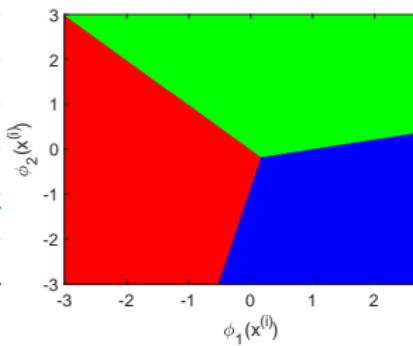


Example: $(\phi_3(x^{(i)}) = 1)$

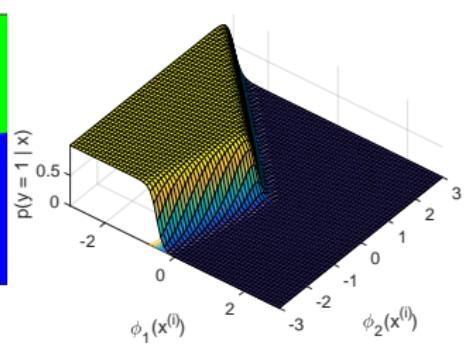
Setup



Decision Boundary

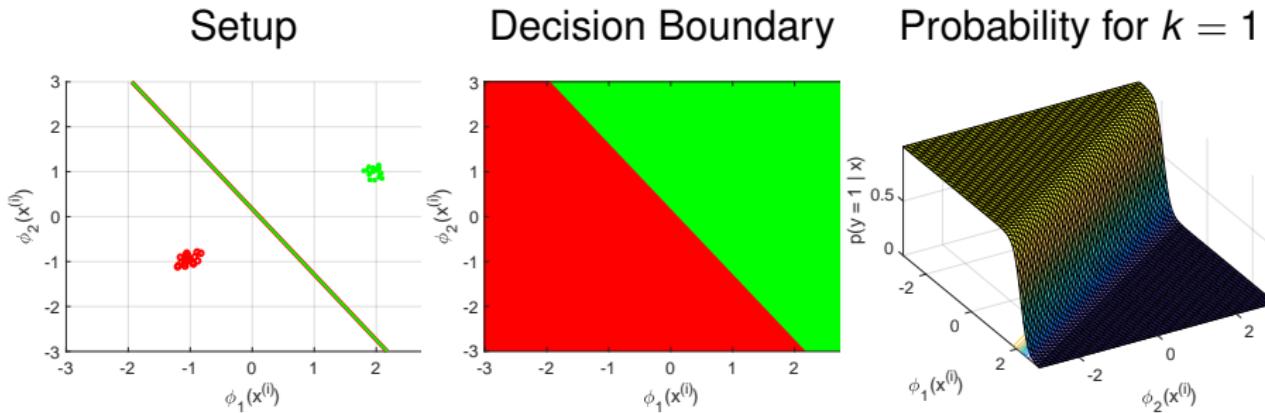


Probability for $k = 1$



Example: What happens when we have two classes?

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

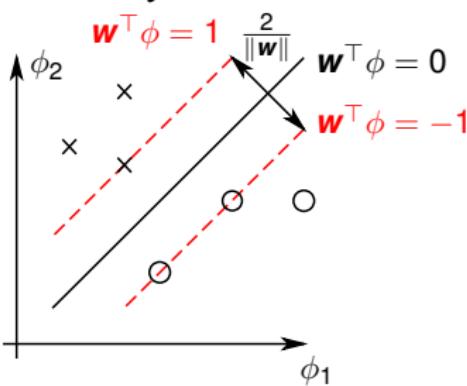


Lines coincide ($\mathbf{w}_{(1)} = -\mathbf{w}_{(2)}$)

How about multi-class SVM?

Recap: Binary SVM

Intuitively:



$$u_1 + c\mathbf{w} = u_2 \quad \mathbf{w}^\top u_2 = -1$$

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

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

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

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

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

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

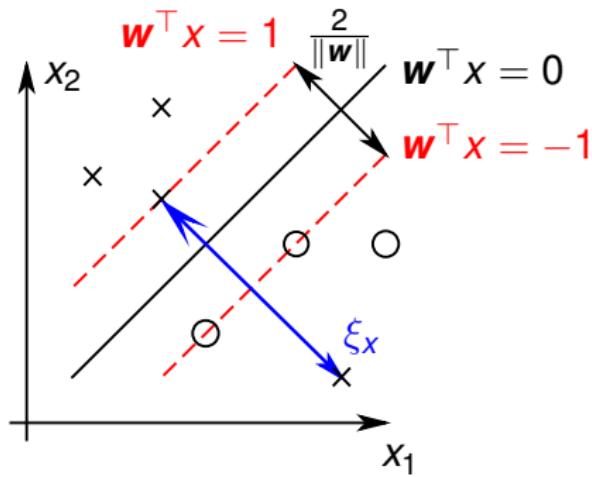
Issue: what if data not linearly separable?

Recap: Binary SVM

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

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

Intuitively:



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

What do we want?

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

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

Multiclass SVM objective:

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

Define mapping:

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

Equivalent formulation:

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

Multiclass SVM objective:

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

Let's get rid of the slack variable:

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

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

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

How does multiclass SVM relate to multiclass logistic regression?

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

Same temperature argument as in the last lecture.

Homework

Solution:

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

Our current framework:

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

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:

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

Prediction with dual variables:

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

Dual of SVM:

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

Prediction with dual variables:

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

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:

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

Example:

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

How to test validity without explicitly constructing feature vectors:

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

Example kernels:

- 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 ($c > 0$):

- 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:

- 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