CS 446/ECE 449: Machine Learning

A. G. Schwing

University of Illinois at Urbana-Champaign, 2020

Scribe & Exercises

L6: Support Vector Machines

Goals of this lecture

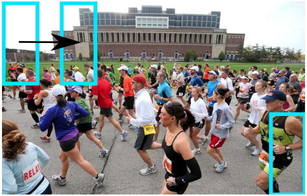
- Getting to know binary Support Vector Machines (SVMs)
- Understanding the relation between SVMs and logistic regression
- Practicing duality

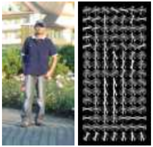
Reading material

- K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 14.5

Sliding window based object detection:

- Scan image at different scales and locations (bounding box size remains identical)
- Extract features for each bounding box (HOG: histogram of oriented gradients)
- Run SVM classifier on bounding box features





2D viz

2D viz

To train SVM we create a dataset $\mathcal{D} = \{(\phi(x^{(i)}), y^{(i)})\}$:

- Bounding box label $y^{(i)} \in \{-1, 1\}$
- Bounding box feature $\phi(x^{(i)})$

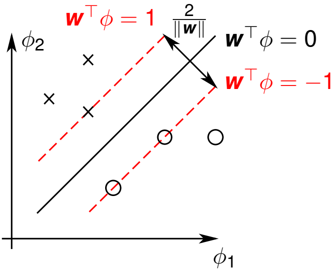
Positive examples:

Negative examples:

Note: Logistic regression would be equally applicable but SVM was the dominant approach.

Binary SVM

Intuitively:



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

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

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

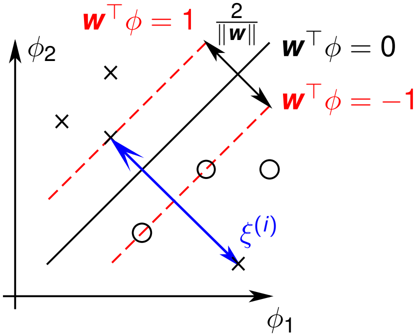
Issue: what if data not linearly separable?

Binary SVM

Introduce slack variables ξ:

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

Intuitively:



Binary SVM:

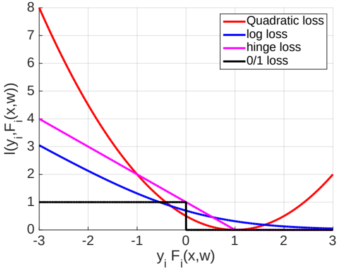
$$\min_{\mathbf{w},\xi^{(i)}\geq 0} \frac{C}{2}\|\mathbf{w}\|_2^2 + \sum_{i\in\mathcal{D}} \xi^{(i)} \qquad \text{s.t.} \qquad y^{(i)}\mathbf{w}^\top\phi(x^{(i)}) \geq 1 - \xi^{(i)} \quad \forall i \in \mathcal{D}$$

Equivalent:

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

Empirical risk minimization:

$$\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{i\in\mathcal{D}} \ell(y^{(i)}, F(x^{(i)}, \mathbf{w}))$$



How to optimize the binary SVM objective (primal problem):

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

Approaches:

- Optimize the primal via gradient descent
- Optimize the corresponding dual problem

Optimization in the primal:

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

What is the gradient of $\max\{0, x\}$?

$$\delta(x \geq 0) = \left\{ \begin{array}{ll} 0 & \text{if } x < 0 \\ 1 & \text{otherwise} \end{array} \right.$$

Keep this in mind for max-pooling.

Optimization in the dual:
Primal objective with constraints:

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

How to obtain the dual objective?
Dual variables $\alpha^{(i)} \geq 0$ for each inequality constraint
Lagrangian:

$$\begin{aligned} L(\cdot) &= \frac{C}{2}\|\mathbf{w}\|_2^2 + \sum_i \xi^{(i)} + \sum_i \alpha^{(i)}(1 - \xi^{(i)} - y^{(i)}\mathbf{w}^T\phi(x^{(i)})) \\ &= \frac{C}{2}\|\mathbf{w}\|_2^2 - \mathbf{w}^T\sum_i \alpha^{(i)}y^{(i)}\phi(x^{(i)}) + \sum_i \xi^{(i)}(1-\alpha^{(i)}) + \sum_i \alpha^{(i)} \end{aligned}$$

Lagrangian:

$$L(\cdot) = \frac{C}{2}\|\boldsymbol{w}\|_2^2 - \boldsymbol{w}^T \sum_i \alpha^{(i)} y^{(i)} \phi(x^{(i)}) + \sum_i \xi^{(i)}(1 - \alpha^{(i)}) + \sum_i \alpha^{(i)}$$

How to obtain the dual program? Optimize the Lagrangian w.r.t. primal variables

- W.r.t. parameters $\boldsymbol{w}$:

$$\frac{\partial L}{\partial \boldsymbol{w}} : \qquad C\boldsymbol{w} = \sum_i \alpha^{(i)} y^{(i)} \phi(x^{(i)})$$

- W.r.t. slack variables $\xi^{(i)}$:

$$\min_{\xi^{(i)} \geq 0} \ \xi^{(i)}(1 - \alpha^{(i)}) \qquad \Longrightarrow \qquad \alpha^{(i)} \leq 1$$

Dual program:

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

Dual program:

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

The dual is quadratic, hence QP solvers are directly applicable.
Techniques such as 'sequential minimal optimization' (J. Platt 1998) are useful.

Recap:

- Linear regression:

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

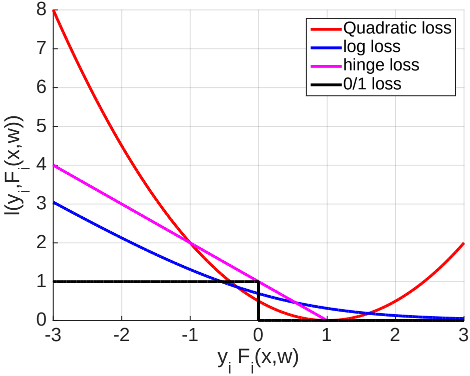
- Logistic regression:

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

- 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
l(y_i,F_i(x,w))
y_i F_i(x,w)
8
7
6
5
4
3
2
1
0
-3
-2
-1
0
1
2
3

Other loss functions:

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

Combining log- and hinge-loss:

$$
\begin{aligned}
\lim_{\epsilon\to 0} \epsilon \log\left(1 + \exp\frac{-F}{\epsilon}\right) \quad &= \\
&\overset{F\geq 0}{=} \quad 0 \\
&\overset{F\leq 0}{=} \quad \lim_{\epsilon\to 0} \frac{\log\left(1+\exp\frac{-F}{\epsilon}\right)}{1/\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{\frac{\exp\frac{-F}{\epsilon}}{1+\exp\frac{-F}{\epsilon}} \cdot (F/\cancel{\epsilon^2})}{-1/\cancel{\epsilon^2}} \\
&= \lim_{\epsilon\to 0} \frac{1}{1+\exp\frac{F}{\epsilon}} \cdot (-F) \\
&= -F
\end{aligned}
$$

0

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:

- 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:

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

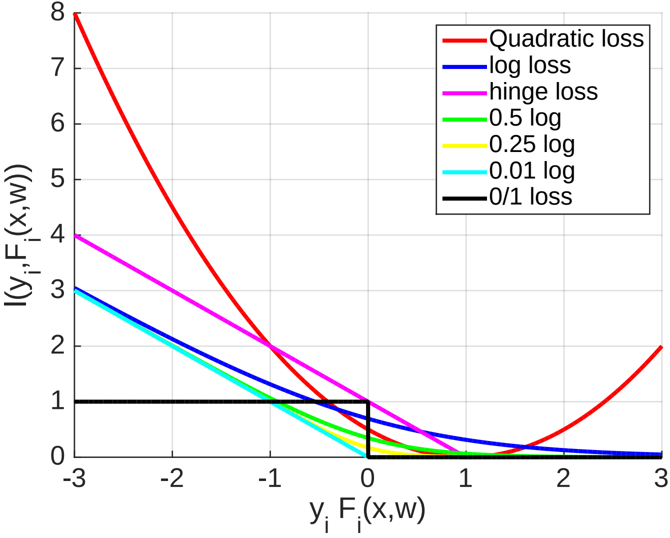
- Binary SVM:

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

- 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



Quiz:

- What are convenient properties of the SVM dual program?
- Relationship between logistic regression and binary SVM?
- How to extend all discussed formulations to more than two classes?

Important topics of this lecture

- Object detection method
- SVMs
- Relationship to linear and logistic regression
- Practicing duality

Up next:

- Other types of features $\phi(x^{(i)})$