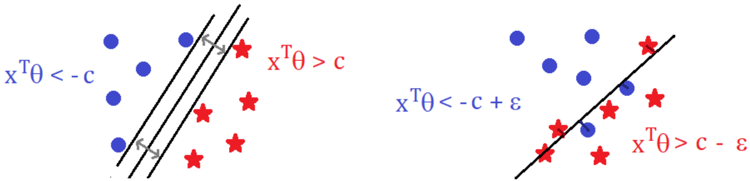
6. Max Margin Classifier

- Margin
- Logistic regression revisited
- Generalized margin classifiers
- Support vector machines
- Soft margins

Classification margins



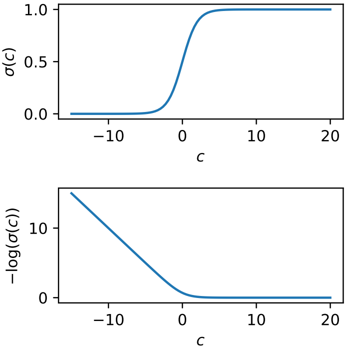
- The bigger the margin c, the better the classifier
- Not all datasets are separable $\rightarrow$ soft margins

Logistic regression

- Data: $x_1, ..., x_m$
- Labels: $y_1, ..., y_m \in \{-1, 1\}$
- Training

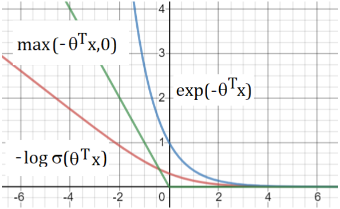
$$\theta^* = \underset{\theta}{\text{argmax}} \sum_{i=1}^{m} \overbrace{\log(\sigma(\underbrace{y_i x_i^T \theta}_{\text{margin}}))}^{\text{Monotonic penalty}}$$

- Classifier: $y(x) = \mathbf{sign}(x^T \theta^*)$



General margin maximizing method

- Data: $x_1, \ldots, x_m$
- Labels: $y_1, \ldots, y_m \in \{-1, 1\}$
- Monotonically increasing penalty g
  - Hinge loss: $g(\xi) = \max(\xi, 0)$
  - Exponential loss: $g(\xi) = e^{-\xi}$
  - Logistic loss: $g(\xi) = -\log(\sigma(\xi))$



- Training

$$\theta^* = \operatorname*{argmax}_{\theta} \sum_{i=1}^{m} g(y_i x_i^T \theta)$$

- Classifier: $y(x) = \mathbf{sign}(x^T \theta^*)$

Is Ridge regression a margin maximizing method?

$$\underset{\theta}{\text{minimize}} \quad \|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2$$

Well... if $y_i \in \{-1, 1\}$, then

$$(\theta^T x_j - y_j)^2 = (1 - \underbrace{y_j \theta^T x_j}_{\text{margin}})^2$$

- Ridge regression tries to force margin to be 1
- It is not margin maximizing, since it does not promote margin to be > 1

Optimum of logistic regression

- Question: What is the global optimum for logistic regression?

$$\theta^* = \operatorname*{argmax}_\theta \underbrace{\sum_{i=1}^{m} \log(\sigma(y_i x_i^T \theta))}_{f(\theta)}$$

Optimum of logistic regression

- Question: What is the global optimum for logistic regression?

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \underbrace{\sum_{i=1}^{m} \log(\sigma(y_i x_i^T \theta))}_{f(\theta)}$$

- Ans: take gradient, set to 0

$$\nabla f(\theta) = A^T d, \qquad A = \begin{bmatrix} x_1^T y_1 \\ \vdots \\ x_m^T y_m \end{bmatrix}, \qquad d_i = 1 - \sigma(y_i x_i^T \theta)$$

Optimum of logistic regression

- Question: What is the global optimum for logistic regression?

$$\theta^* = \underset{\theta}{\text{argmax}} \underbrace{\sum_{i=1}^{m} \log(\sigma(y_i x_i^T \theta))}_{f(\theta)}$$

- Ans: take gradient, set to 0

$$\nabla f(\theta) = A^T d, \qquad A = \begin{bmatrix} x_1^T y_1 \\ \vdots \\ x_m^T y_m \end{bmatrix}, \qquad d_i = 1 - \sigma(y_i x_i^T \theta)$$

- When does $\nabla f(\theta) = 0$?

Optimum of logistic regression

- Question: What is the global optimum for logistic regression?

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \underbrace{\sum_{i=1}^{m} \log(\sigma(y_i x_i^T \theta))}_{f(\theta)}$$

- Ans: take gradient, set to 0

$$\nabla f(\theta) = A^T d, \qquad A = \begin{bmatrix} x_1^T y_1 \\ \vdots \\ x_m^T y_m \end{bmatrix}, \qquad d_i = 1 - \sigma(y_i x_i^T \theta)$$

- When does $\nabla f(\theta) = 0$? Ans: assuming perfect classification ($y_i x_i^T \theta > 0$ for all i), take $\|\theta\|_2 \to +\infty$

- But since $y = \mathbf{sign}(x^T \theta)$, scaling on θ doesn't matter

Hard margin support vector machines

- Margin c of dataset

$$y_i x_i^T \theta \geq c \, \forall i \overset{\text{normalize } \theta}{\Longleftrightarrow} c = \frac{1}{\|\theta\|_2}, \; y_i x_i^T \theta > 1$$

- Hard margin support vector machine (SVM):

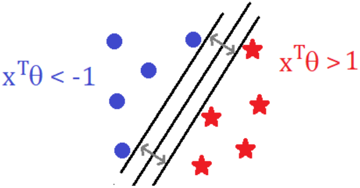
$$\underset{\theta}{\text{minimize}} \quad \|\theta\|_2^2 \qquad \text{subject to} \quad y_i x_i^T \theta > 1 \, \forall i$$

- Equivalent to maximizing margin $\frac{1}{\|\theta\|_2}$
- How to solve?
  - Quadratic programming interior point solver. (Super expensive)
  - Projected gradient descent. (How to project on halfspaces?)

Hard margin support vector machine

- Data: $x_1, ..., x_m$
- Labels: $y_1, ..., y_m \in \{-1, 1\}$
- Training θ^* optimizes

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \|\theta\|_2^2 \\ \text{subject to} & y_i x_i^T \theta \geq 1, \forall i \end{array}$$

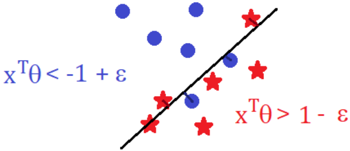


- Classifier: $y(x) = \mathbf{sign}(x^T \theta^*)$

Soft margin support vector machine

- Data: $x_1, ..., x_m$
- Labels: $y_1, ..., y_m \in \{-1, 1\}$
- Training θ^* optimizes

$$\begin{array}{ll} \underset{\theta,\xi}{\text{minimize}} & \|\theta\|_2^2 + \lambda \sum_i \xi_i \\ \text{subject to} & y_i x_i^T \theta \geq 1 + \xi_i, \forall i \\ & \xi_i \geq 0,\ \forall i \end{array}$$



- Classifier: $y(x) = \mathbf{sign}(x^T\theta^*)$

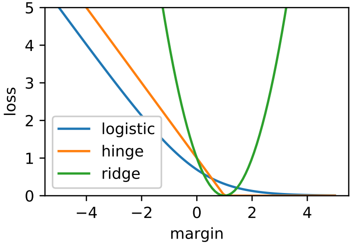
Summary of margin methods

SVM (soft margin)

$$\min \theta \; \frac{1}{2}\|\theta\|_2^2 + \lambda \sum_{j=1}^{m} \max\{1 - y_j \theta^T x_j, 0\}$$

Ridge regression

$$\begin{aligned} \min \theta \qquad & \lambda\|\theta\|_2^2 + \sum_{j=1}^{m} (\theta^T x_j - y_j)^2 \\ = \quad & \lambda\|\theta\|_2^2 + (1 - y_j \theta^T x_j)^2 \end{aligned}$$



(Regularized) logistic regression

$$\min_{\theta} \; \lambda\|\theta\|_2^2 + \sum_{j=1}^{m} \log(\sigma(y_j \theta^T x_j)$$

How to solve?

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \; \frac{1}{2}\|\theta\|_2^2 + \lambda \sum_{j=1}^{m} \max\{1 - y_j \theta^T x_j, 0\}$$

- Subgradient descent?

$$g(\theta) = \max\{1 - y_j x_j^T \theta, 0\}, \quad \partial g(\theta) = ?$$

How to solve?

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\theta\|_2^2 + \lambda \sum_{j=1}^{m} \max\{1 - y_j \theta^T x_j, 0\}$$

- Subgradient descent?

$$g(\theta) = \max\{1 - y_j x_j^T \theta, 0\}, \quad \partial g(\theta) = \begin{cases} \{-y_j x_j\} & y_j x_j^T \theta < 1 \\ \{0\} & y_j x_j^T \theta > 1 \\ y_j x_j \cdot [-1, 0] & y_j x_j^T \theta = 1 \end{cases}$$

How to solve?

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \; \frac{1}{2}\|\theta\|_2^2 + \lambda \sum_{j=1}^{m} \max\{1 - y_j \theta^T x_j, 0\}$$

- Subgradient descent?

$$g(\theta) = \max\{1 - y_j x_j^T \theta, 0\}, \quad \partial g(\theta) =$$

Require diminishing step size, converges slowly

How to solve?

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\theta\|_2^2 + \lambda \sum_{j=1}^{m} \max\{1 - y_j \theta^T x_j, 0\}$$

- Subgradient descent? Require diminishing step size, converges slowly
- Generic QP solver? Quadratic program form:

$$\begin{array}{lll} \underset{\theta \in \mathbb{R}^n}{\text{minimize}} & \frac{1}{2}\|\theta\|_2^2 + \lambda \xi^T \mathbf{1} & \text{(Quadratic objective)} \\ \text{subject to} & y_j x_j^T \theta \cancel{+} \xi_j \geq 1 \ \forall j & \text{(Linear inequality constraints)} \end{array}$$

  - Generic solver like cvx or Yalmip uses interior point solvers
  - At each iteration, solves linear system of order $O(m + n)$
  - Overall complexity per iteration?

How to solve?

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\theta\|_2^2 + \lambda \sum_{j=1}^{m} \max\{1 - y_j \theta^T x_j, 0\}$$

- Subgradient descent? Require diminishing step size, converges slowly
- Generic QP solver? Quadratic program form:

$$\begin{array}{lll} \underset{\theta \in \mathbb{R}^n}{\text{minimize}} & \frac{1}{2}\|\theta\|_2^2 + \lambda \xi^T \mathbf{1} & \text{(Quadratic objective)} \\ \text{subject to} & y_j x_j^T \theta \cancel{+} \xi_j \geq 1 \ \forall j & \text{(Linear inequality constraints)} \end{array}$$

  - Generic solver like cvx or Yalmip uses interior point solvers
  - At each iteration, solves linear system of order $O(m+n)$
  - Overall complexity per iteration? $O((m+n)^3)$ (with no extra tricks)

How to solve?

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \; \frac{1}{2}\|\theta\|_2^2 + \lambda \sum_{j=1}^{m} \max\{1 - y_j \theta^T x_j, 0\}$$

- Subgradient descent? Require diminishing step size, converges slowly
- Generic QP solver? Doesn't scale well with large m, n

How to solve?

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}\|\theta\|_2^2 + \lambda \sum_{j=1}^{m} \max\{1 - y_j \theta^T x_j, 0\}$$

- Subgradient descent? Require diminishing step size, converges slowly
- Generic QP solver? Doesn't scale well with large m, n
- In practice?

How to solve?

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \; \frac{1}{2}\|\theta\|_2^2 + \lambda \sum_{j=1}^{m} \max\{1 - y_j \theta^T x_j, 0\}$$

- Subgradient descent? Require diminishing step size, converges slowly
- Generic QP solver? Doesn't scale well with large m, n
- In practice? Solve the dual form (stay tuned)

Summary

- For $y_i \in \{-1, 1\}$, the classification margin of training sample i is $y_i x_i^T \theta$

$$y_i x_i^T \theta \text{ is } \begin{cases} \text{large, positive} & \rightarrow \text{ well done! good separation} \\ \text{small, positive} & \rightarrow \text{ barely made the cut} \\ \text{negative} & \rightarrow \text{ classified wrong} \end{cases}$$

- Logistic regression is a margin maximizing technique
  - Also, hinge loss, exponential loss
- Support vector machines (SVM): the ultimate margin maximizing framework
- Soft margins: add penalties

$$\xi_i = \max\{-y_i x_i^T \theta, 0\}$$

- How to solve SVM?
  - This form is hard to solve. Stay tuned for dual form.