Convex programs and duality

Dual form of the Perceptron solution

Given a training set of points $\{(x^{(i)}, y^{(i)}) : i = 1 \ldots n\}$:

Perceptron algorithm

- Initialize $w = 0$ and $b = 0$
- While some training point (x, y) is misclassified:
  - $w = w + yx$
  - $b = b + y$

Dual form of the Perceptron solution

Given a training set of points $\{(x^{(i)}, y^{(i)}) : i = 1 \ldots n\}$:

Perceptron algorithm

- Initialize $w = 0$ and $b = 0$
- While some training point (x, y) is misclassified:
  - $w = w + yx$
  - $b = b + y$

The final answer is of the form:

$$w = \sum_i \alpha_i y^{(i)} x^{(i)},$$

where $\alpha_i = \#$ of times an update occurred on point i.

Dual form of the Perceptron solution

Given a training set of points $\{(x^{(i)}, y^{(i)}) : i = 1 \ldots n\}$:

Perceptron algorithm

- Initialize $w = 0$ and $b = 0$
- While some training point (x, y) is misclassified:
  - $w = w + yx$
  - $b = b + y$

The final answer is of the form:

$$w = \sum_i \alpha_i y^{(i)} x^{(i)},$$

where $\alpha_i = \#$ of times an update occurred on point i.

Can equivalently represent w by $\alpha = (\alpha_1, \ldots, \alpha_n)$.

Dual form of the Perceptron algorithm

Perceptron algorithm: primal form

- Initialize $w = 0$ and $b = 0$
- While some training point $(x^{(i)}, y^{(i)})$ is misclassified:
  - $w = w + y^{(i)} x^{(i)}$
  - $b = b + y^{(i)}$

Perceptron algorithm: dual form

- Initialize $w = 0$ and $b = 0$
- While some training point $(x^{(i)}, y^{(i)})$ is misclassified:
  - $\alpha_i = \alpha_i + 1$
  - $b = b + y^{(i)}$

Answer: $w = \sum_i \alpha_i y^{(i)} x^{(i)}$

Maximum-margin linear classifier

- Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$

$$\begin{array}{ll} (\text{PRIMAL}) & \min\limits_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 \\ \text{s.t.:} & y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \ldots, n \end{array}$$

- This is a **convex optimization problem**:
  - Convex objective function
  - Linear constraints

Maximum-margin linear classifier

- Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$

$$\begin{array}{ll} (\text{PRIMAL}) & \min\limits_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 \\ \text{s.t.:} & y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \ldots, n \end{array}$$

- This is a **convex optimization problem**:
  - Convex objective function
  - Linear constraints
- As such, it has a **dual maximization problem**.
- The **primal** and **dual** problems have the same optimum value.

The dual program

$$\text{(PRIMAL)} \qquad \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \ldots, n$$

$$\text{(DUAL)} \qquad \max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)} \cdot x^{(j)})$$

$$\text{s.t.:} \quad \sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

$$\alpha \geq 0$$

The dual program

$$\begin{array}{ll} (\text{PRIMAL}) & \min\limits_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad \frac{1}{2}\|w\|^2 \\ \text{s.t.:} & y^{(i)}(w\cdot x^{(i)}+b)\geq 1 \quad \text{for all } i=1,2,\ldots,n \end{array}$$

$$\begin{array}{ll} (\text{DUAL}) & \max\limits_{\alpha\in\mathbb{R}^n} \quad \sum\limits_{i=1}^{n}\alpha_i - \frac{1}{2}\sum\limits_{i,j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}(x^{(i)}\cdot x^{(j)}) \\ & \text{s.t.:} \quad \sum\limits_{i=1}^{n}\alpha_i y^{(i)} = 0 \\ & \qquad \alpha \geq 0 \end{array}$$

Complementary slackness: At optimality, $w = \sum_{i=1}^{n}\alpha_i y^{(i)}x^{(i)}$ and

$$\alpha_i > 0 \quad \Rightarrow \quad y^{(i)}(w\cdot x^{(i)}+b) = 1$$

Points $x^{(i)}$ with $\alpha_i > 0$ are **support vectors**.

Dual of soft-margin SVM

$$\text{(PRIMAL)} \qquad \min_{w\in\mathbb{R}^d, b\in\mathbb{R}, \xi\in\mathbb{R}^n} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.:} \ \ y^{(i)}(w\cdot x^{(i)} + b) \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n$$

$$\xi \geq 0$$

$$\text{(DUAL)} \qquad \max_{\alpha\in\mathbb{R}^n} \ \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}(x^{(i)}\cdot x^{(j)})$$

$$\text{s.t.:} \ \ \sum_{i=1}^{n}\alpha_i y^{(i)} = 0$$

$$0 \leq \alpha_i \leq C$$

At optimality, $w = \sum_i \alpha_i y^{(i)} x^{(i)}$, with

$$0 < \alpha_i < C \quad \Rightarrow \quad y^{(i)}(w\cdot x^{(i)} + b) = 1$$

$$\alpha_i = C \quad \Rightarrow \quad y^{(i)}(w\cdot x^{(i)} + b) = 1 - \xi_i$$

A high-level view of optimization

Unconstrained optimization

Logistic regression: find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that minimize

$$L(w, b) = \sum_{i=1}^{n} \ln(1 + \exp(-y^{(i)}(w \cdot x^{(i)} + b)).$$

Can check for convexity and solve using gradient descent or SGD.

A high-level view of optimization

Unconstrained optimization

Logistic regression: find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that minimize

$$L(w, b) = \sum_{i=1}^{n} \ln(1 + \exp(-y^{(i)}(w \cdot x^{(i)} + b)).$$

Can check for convexity and solve using gradient descent or SGD.

Constrained optimization

Support vector machine: find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that minimize

$$L(w) = \|w\|^2$$

subject to the constraints

$$y^{(i)}(w \cdot x^{(i)} + b) \geq 1$$

What problems of this kind are easy to solve?

Constrained optimization

Write the optimization problem in a standardized form:

$$\min f_o(z)$$
$$f_i(z) \leq 0 \quad \text{for } i = 1, .., m$$
$$h_i(z) = 0 \quad \text{for } i = 1, .., n$$

Special cases that can be solved (relatively) easily:

- **Linear programs**:
  f_o, f_i, h_i are all linear functions.
- **Convex programs**:
  f_o, f_i are convex functions. The h_i are linear functions.

Convex surrogates for 0-1 loss

Want a separator w that misclassifies as few training points as possible.

- 0-1 loss: charge 1 for each misclassified (x, y)

But this leads to an NP-hard optimization problem!

Convex surrogates for 0-1 loss

Want a separator w that misclassifies as few training points as possible.

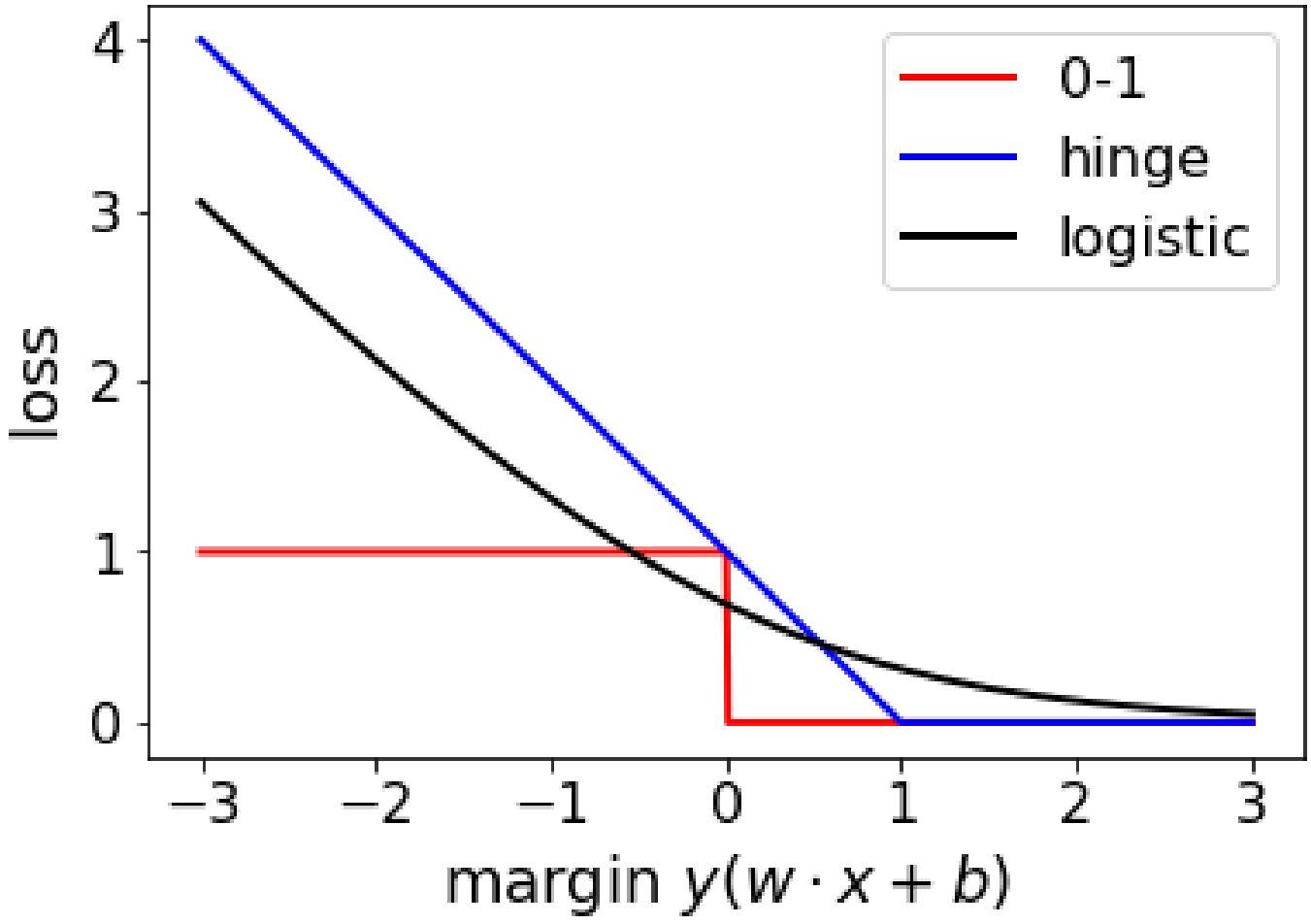
- 0-1 loss: charge 1 for each misclassified (x, y)

But this leads to an NP-hard optimization problem!

Instead, use **convex** loss functions.

- Hinge loss (SVM): charge

$$\begin{array}{cc} 0 & \text{if } y(w \cdot x + b) \geq 1 \\ 1 - y(w \cdot x + b) & \text{otherwise} \end{array}$$

- Logistic loss: charge $\ln(1 + e^{-y(w \cdot x + b)})$

Convex loss functions