Duality and convex programs

CSE 250B

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

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

Hard-margin SVM

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

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

- 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

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

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

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

$$\begin{array}{ll} \text{(PRIMAL)} & \min_{w\in\mathbb{R}^d, b\in\mathbb{R}, \xi\in\mathbb{R}^n} \|w\|^2 + C\sum_{i=1}^n \xi_i \\ \text{s.t.:} & y^{(i)}(w\cdot x^{(i)}+b)\geq 1-\xi_i \quad \text{for all } i=1,2,\ldots,n \\ & \xi \geq 0 \end{array}$$

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

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.

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:

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

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.

Example: regression with ℓ_1 loss

Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \mathbb{R}$, find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ minimizing

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

Optimization over $d + 1 + n$ variables, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ and $z \in \mathbb{R}^n$:

$$\begin{aligned} \min \sum_{i=1}^{n} z_i & \\ y^{(i)} - w \cdot x^{(i)} - b &\leq z_i, \quad i = 1, 2, \ldots, n \\ w \cdot x^{(i)} + b - y^{(i)} &\leq z_i, \quad i = 1, 2, \ldots, n \end{aligned}$$

A linear program.

The dual of an optimization problem

Primal optimization problem (may or may not be convex):

$$\begin{aligned} \min_{z \in \mathbb{R}^d} & f_o(z) \\ f_i(z) &\leq 0 \quad \text{for } i = 1, .., n \\ h_i(z) &= 0 \quad \text{for } i = 1, .., m \end{aligned}$$

- **Feasible region**: $\Omega = \{z \in \mathbb{R}^d : f_i(z) \leq 0, h_i(z) = 0, \text{for all } i\}$.
- Optimal solution z^*, with value p^*.

Create a related unconstrained problem:

- The **Lagrangian**: for $\lambda = (\lambda_1, \ldots, \lambda_n)$ and $\nu = (\nu_1, \ldots, \nu_m)$,

$$L(z, \lambda, \nu) = f_o(z) + \sum_{i=1}^{n} \lambda_i f_i(z) + \sum_{i=1}^{m} \nu_i h_i(z).$$

- Can minimize for any given λ, ν:

$$g(\lambda, \nu) = \inf_{z \in \mathbb{R}^d} L(z, \lambda, \nu).$$

Lagrangian: a lower bound on the optimum value

- Primal optimal solution $z^* \in \Omega$ has value $p^* = f_o(z^*)$.
- Lagrangian: $L(z, \lambda, \nu) = f_o(z) + \sum_{i=1}^{n} \lambda_i f_i(z) + \sum_{i=1}^{m} \nu_i h_i(z)$.
- $g(\lambda, \nu) = \inf_{z \in \mathbb{R}^d} L(z, \lambda, \nu)$.

Claim 1: $g(\lambda, \nu) \leq p^*$ for any $\lambda \geq 0$ and any ν.

Lagrangian: a lower bound on the optimum value

- Primal optimal solution $z^* \in \Omega$ has value $p^* = f_o(z^*)$.
- Lagrangian: $L(z, \lambda, \nu) = f_o(z) + \sum_{i=1}^{n} \lambda_i f_i(z) + \sum_{i=1}^{m} \nu_i h_i(z)$.
- $g(\lambda, \nu) = \inf_{z \in \mathbb{R}^d} L(z, \lambda, \nu)$.
- $g(\lambda, \nu) \leq p^*$ for any $\lambda \geq 0$ and any ν.

Claim 2: $g(\lambda, \nu)$ is a concave function (even if the primal isn't convex).

The dual optimization problem

Primal optimization problem (may or may not be convex):

$$\min_{z \in \mathbb{R}^d} f_o(z)$$
$$f_i(z) \leq 0 \quad \text{for } i = 1, .., n$$
$$h_i(z) = 0 \quad \text{for } i = 1, .., m$$

Feasible region Ω, optimum $z^* \in \Omega$ with value $p^* = f_o(z^*)$.

- Lagrangian: $L(z, \lambda, \nu) = f_o(z) + \sum_{i=1}^n \lambda_i f_i(z) + \sum_{i=1}^m \nu_i h_i(z)$.
- $g(\lambda, \nu) = \inf_{z \in \mathbb{R}^d} L(z, \lambda, \nu)$.

Dual optimization problem over $n + m$ variables:

$$\max g(\lambda, \nu)$$
$$\lambda \geq 0$$

Convex problem with optimal solution λ^*, ν^*, with value d^*.

"Weak duality": $d^* \leq p^*$.

Strong duality

If the primal problem is convex and if *Slater's condition* holds:

- There is some z with $f_i(z) < 0$ for $i = 1, \ldots, n$

then $d^* = p^*$.

- **Certificate of optimality**: if you can exhibit z, λ, ν such that $f_o(z) = g(\lambda, \nu)$, then all must be optimal.
- **Complementary slackness**: under strong duality, the optimal z^*, λ^*, ν^* must satisfy:

$$\lambda_i^* f_i(z^*) = 0 \quad \text{for all } i = 1, \ldots, n$$

Complementary slackness

Under strong duality, let z^* be the primal optimal solution and (λ^*, ν^*) the dual optimal solution. So $f_o(z^*) = g(\lambda^*, \nu^*)$.

Claim: For all i, we have $\lambda_i^* f_i(z^*) = 0$.

Example: Hard-margin SVM

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

$$\text{(PRIMAL)} \qquad \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \|w\|^2$$
$$\text{s.t.:} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \ldots, n$$

Step 1: Convert to standard form

Step 2: Write down the Lagrangian, look at first derivative

Step 3: Write down the dual

Step 4: What does complementary slackness tell us?

In retrospect: convex losses for classification

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

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

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

Convex loss functions