11.. Support vector machines

- Remind) Logistic Regression : classify 2 (or K) class based on decision boundary
  if inner product $\langle\theta, x\rangle$ is larger than threshold.

* Support Vector Machines (SVM)

Goal : 고차원 공간에서 선형 결정 경계를 통해 data를 분류하자

Idea : Margin을 최대화해서 일반화 성능이 가장 좋은 classifier를 찾자.

비슷하게 SVM 에서 decision boundary 학습 but data distribution 가정 안함

11.1 The perceptron algorithm (Limit & Problem)

- 시작점 : Consider binary classification
- Dataset : $D = \{(x_1, y_1), \cdots (x_n, y_n)\} \in \mathbb{R}^d \times \{-1, 1\}$
- Goal : Hyperlane 을 따라 2가지 클래스로 분류하는 classifier를 찾자.
- Classifier f :

$$f(x) = \begin{cases} 1 , & \langle\theta, x\rangle \geq b \\ -1 , & \langle\theta, x\rangle < b \end{cases}$$

* 전제 조건 : data is linearly separable, data가 선형적으로 완전히 분리가능해야만 작동

* 문제점 : 1) 해가 존재하지 않을 수 있음 (linearly non-separable data)

2) 여러 해가 존재하면 어떤 θ를 선택할지 기준 없음 ($\overset{\text{parameter}}{\theta}$ and $\overset{\text{threshold}}{b}$ aren't necessarily unique)

↳ SVM Motivation ① fully linear separable 을 가정 → ② 완벽히 선형분리되지 않는 case (duality 도입)

11.2 Hard-Margin SVM

- Goal : fully linearly separable data 에 대해 Margin을 최대화하는 Hyperplane 찾기

  In more detail : we want to maximize margin $m \geq 0$ s.t.

  1) "+1"로 분류된 모든 points는 Hyperplane의 양수 쪽 (positive side) 에 있고, 그 Hyperplane 까지의 거리는 최소 m.

  1) "-1"로 분류된 모든 points는 Hyperplane의 음수 쪽 (negative side) 에 있고, 그 Hyperplane 까지의 거리는 최소 m.

- distance of a point x to a hyperplane $\mathcal{H} = \{x : \langle\theta, x\rangle = b\}$:

$$\text{distance}(x, H) = \frac{|\langle\theta, x\rangle - b|}{\|\theta\|_2} \qquad \frac{|\langle\theta, x - x_0\rangle|}{\|\theta\|_2} = \frac{|\langle\theta, x\rangle - b|}{\|\theta\|_2} \quad ***$$

By definition, the vector θ is perpendicular to Hyperplane $\mathcal{H}$ ⇒ Any vector lies on Hyperplane.

만약 $x_0, x_1 \in \mathcal{H}$ 일때 $x_1 - x_0$ vector도 $\mathcal{H}$ 위에 있고, orthogonal to θ :

$$\langle x_1 - x_0, \theta\rangle = \langle x_1, \theta\rangle - \langle x_0, \theta\rangle = b - b = 0.$$

⇒ θ가 $\mathcal{H}$에 perpendicular, 최소 거리 ($x \sim \mathcal{H}$)는 projection of $x - x_0$ onto θ, where $x_0 \in \mathcal{H}$.

이걸 수식으로 표현하면 ***

이걸 우리 Goal에 대입하면 : $y_i \dfrac{|\langle\theta, x_i\rangle - b|}{\|\theta\|_2} \geq m.$

- Optimization Task :

1) Margin 포함

$$\max_{m, \theta, b} m \text{ subject to } y_i \frac{|\langle\theta, x_i\rangle - b|}{\|\theta\|_2} \geq m, \quad i \in \{1, \cdots, n\}$$

2) scale normalization

$\|\theta\|_2 = \frac{1}{m}$로 설정하면 → $\displaystyle \max_{\theta, b} \frac{1}{\|\theta\|_2}$ subject to $y_i \dfrac{|\langle\theta, x_i\rangle - b|}{\cancel{\|\theta\|_2}} \geq \dfrac{1}{\cancel{\|\theta\|_2}}$

미분할때 수식 쉽게 만들고, min 으로 하려고... $f(\theta) = \frac{1}{2}\|\theta\|^2 = \frac{1}{2}\theta^T\theta \rightarrow \nabla_\theta f(\theta) = \nabla_\theta \left(\frac{1}{2}\theta^T\theta\right) = \theta$

11.3 Soft-Margin SVM → 현실 data는 대부분 완벽히 선형분리 불가능 → duality 적용

Noise or Overlap이 있는 현실 data를 어떻게 다룰까? → Margin 조건을 일부 위반 허용하자 → slack variables ξ_i

- 새로운 constraint : $y_i(\langle\theta, x_i\rangle - b) \geq 1 - \xi_i \quad (\xi_i \geq 0)$

we take the slack variable in the objective into account by penalizing large values of ξ_i

- 새로운 optimization : $\min_{\theta, b} \frac{1}{2}\|\theta\|^2 + \lambda \sum_{i=1}^{n} \xi_i$ subject to $y_i(\langle\theta, x_i\rangle - b) \geq 1 - \xi_i$, $\xi_i \geq 0$

$$\min_{\theta, b} \frac{1}{2}\|\theta\|^2 + \lambda \sum_{i=1}^{n} \max(1 - y_i(\langle\theta, x_i\rangle - b), 0)$$

($y_i(\langle\theta, x_i\rangle - b) \geq 1 - \xi_i$
$\xi_i \geq 0$ 합침.

$$\min_{\theta, b} \frac{1}{2}\|\theta\|^2 + \lambda \sum_{i=1}^{n} \xi_i \quad \text{subject to} \quad \xi_i = \max(1 - y_i(\langle\theta, x_i\rangle - b), 0)$$