Lecture 12

§1 Derivation Ⅱ: hinge loss

1、利用 indicator function 改写 objective function (以 logistic regression 为例)

- Logistic regression 的 hypothesis function 为

$$f_{\mathbf{w},b}(\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} = g(z)$$

其中 $z = \mathbf{w}^T\mathbf{x}$

- 其 cross-entropy loss 为：

$$\text{cost}(y, f_{\mathbf{w},b}(\mathbf{x})) = \begin{cases} -\log(f_{\mathbf{w},b}(\mathbf{x})) & , \text{if } y=1 \\ -\log(1-f_{\mathbf{w},b}(\mathbf{x})) & , \text{if } y=-1 \end{cases}$$

- 利用 indicator function，可化为

$$\text{cost}(y, f_{\mathbf{w},b}(\mathbf{x})) = -\delta_{y=1}\log(f_{\mathbf{w},b}(\mathbf{x})) - \delta_{y=-1}\log(1-f_{\mathbf{w},b}(\mathbf{x}))$$

其中 $\delta_a = 1$ 若 a 为 true，否则为 0

- 因此，the objective function of the regularized logistic regression 为：

$$J(\mathbf{w}) = -\frac{1}{m}\sum_{i=1}^{m}[\delta_{y=1}\log(f_{\mathbf{w},b}(\mathbf{x}_i)) + \delta_{y=-1}\log(1-f_{\mathbf{w},b}(\mathbf{x}_i))] + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

2、利用 hinge loss 表示 SVM 的 objective function

- SVM 的 objective function 为

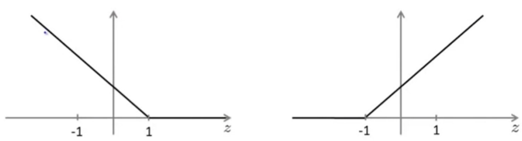
$$\frac{1}{m}\sum_{i=1}^{m}[\delta_{y_i=1}\,\text{cost}_1(\mathbf{w}^T\mathbf{x}_i+b) + \delta_{y_i=-1}\,\text{cost}_{-1}(\mathbf{w}^T\mathbf{x}_i+b)] + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

$$= C\sum_{i=1}^{m}[\delta_{y_i=1}\,\text{cost}_1(\mathbf{w}^T\mathbf{x}_i+b) + \delta_{y_i=-1}\,\text{cost}_{-1}(\mathbf{w}^T\mathbf{x}_i+b)] + \frac{1}{2}\sum_{j=1}^{n} w_j^2$$

其中 $C = \frac{1}{\lambda}$

- cost 的选取

Hinge loss: $\max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))$



- If $y_i = +1$, we require that $\mathbf{w}^\top\mathbf{x}_i + b \geq 1$. In other words, $\text{cost}_1(\mathbf{w}^\top\mathbf{x}_i + b) = 0$ if $\mathbf{w}^\top\mathbf{x}_i + b \geq 1$
- If $y_i = -1$, we require that $\mathbf{w}^\top\mathbf{x}_i + b \leq -1$. In other words, $\text{cost}_{-1}(\mathbf{w}^\top\mathbf{x}_i + b) = 0$ if $\mathbf{w}^\top\mathbf{x}_i + b \leq -1$

- Mathematics behind hinge loss

- However, hinge loss is non-smooth. We transform the objective function of support vector machine to the following

希望使得 hinge loss 为 0

$$\min_{\mathbf{w},b} \frac{1}{2}\sum_{j=1}^{n} w_j^2 \tag{3}$$

$$s.t.\ \mathbf{w}^\top\mathbf{x}_i + b \geq 1, \text{if } y_i = 1;\ \ \mathbf{w}^\top\mathbf{x}_i + b < -1, \text{if } y_i = -1.$$

- It can be simplified as follows

hinge loss 的 objective function

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{4}$$

$$s.t.\ y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1, \forall i$$

- Utilizing $p_i = \frac{\mathbf{w}^\top \mathbf{x}_i + b}{\|\mathbf{w}\|}$, which denotes the projection length of $\mathbf{x}_i$ on $\mathbf{w}$ or the distance from $\mathbf{x}_i$ to the decision boundary $\mathbf{w}^\top \mathbf{x} + b = 0$, we have

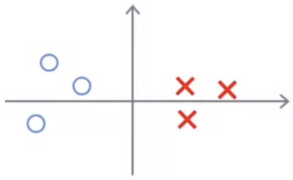
$$\mathbf{w}^\top \mathbf{x}_i + b = p_i \cdot \|\mathbf{w}\| \tag{5}$$

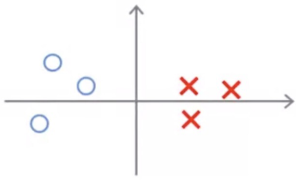
- The objective function of support vector machine is transformed to

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{6}$$
$$s.t.\ y_i \cdot p_i \cdot \|\mathbf{w}\| \geq 1, \forall i$$

与large margin结论相同

- Let's see the following two decision boundaries (plot below)
- If the projection length p_i is larger, then $\|\mathbf{w}\|$ could be smaller, leading to better solution. Thus, **we prefer large margin**.





§2 KKT conditions

Lagrange duality

- Given a general minimization problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_i(\mathbf{x}) \leq 0, \quad i = 1,\dots,m \\ & \ell_j(\mathbf{x}) = 0, \quad j = 1,\dots,r \end{aligned}$$

Note that here $\mathbf{x}$ denotes the argument we aim to optimize, rather than a data point.
- The **Lagrangian function**:

$$L(\mathbf{x},\mathbf{u},\mathbf{v}) = f(\mathbf{x}) + \sum_{i=1}^{m} u_i h_i(\mathbf{x}) + \sum_{j=1}^{r} v_j \ell_j(\mathbf{x})$$

- The **Lagrange dual function**:

$$g(\mathbf{u},\mathbf{v}) = \min_{\mathbf{x}\in\mathbb{R}^n} L(\mathbf{x},\mathbf{u},\mathbf{v})$$

- The **dual problem**:

$$\max_{\mathbf{u}\in\mathbb{R}^m,\mathbf{v}\in\mathbb{R}^r} g(\mathbf{u},\mathbf{v})$$
$$\text{subject to } \mathbf{u} \geq 0$$

KKT conditions

- Given general problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_i(\mathbf{x}) \leq 0, \quad i = 1,\dots,m \\ & \ell_j(\mathbf{x}) = 0, \quad j = 1,\dots,r \end{aligned}$$

- The **Karush-Kuhn-Tucker conditions** or **KKT conditions** are:
  - $0 \in \partial f(\mathbf{x}) + \sum_{i=1}^{m} u_i \partial h_i(\mathbf{x}) + \sum_{j=1}^{r} v_j \partial \ell_j(\mathbf{x})$ (stationarity)
  - $u_i \cdot h_i(\mathbf{x}) = 0$ for all i (complementary slackness)
  - $h_i(\mathbf{x}) \leq 0, \ell_j(\mathbf{x}) = 0$ for all i, j (primal feasibility)
  - $u_i \geq 0$ for all i (dual feasibility)

接下来考虑用KKT解SVM问题(数据点能完全分隔的情况)

§3 Optimization of SVM (完全线性可分)

1. Optimization of SVM ('w')

- 根据先前分析，SVM的objective function为

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i+b)\geq 1,\quad \forall i$$

其可以转化为

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{s.t. } 1-y_i(\mathbf{w}^T\mathbf{x}_i+b)\leq 0$$

其 Lagrange function 为

$$L(\mathbf{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\mathbf{w}\|^2+\sum_i^m \alpha_i(1-y_i(\mathbf{w}^T\mathbf{x}_i+b))\ (\alpha_i\text{为添加变量})\quad \alpha_i\geq 0\ \ \forall i$$

注：我们希望

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{w},b} L(\mathbf{w},b,\boldsymbol{\alpha})$$
$$\text{s.t. } \boldsymbol{\alpha}\geq 0$$

- 因此 primal 与 dual optimal solution 需满足以下 KKT conditions：

① Stationarity：

$$\frac{\partial L}{\partial \mathbf{w}}=0 \Rightarrow \mathbf{w}-\sum_i^m \alpha_i y_i \mathbf{x}_i=0$$
$$\Rightarrow \mathbf{w}=\sum_i^m \alpha_i y_i \mathbf{x}_i \quad ①$$
$$\frac{\partial L}{\partial b}=0 \Rightarrow \sum_i^m \alpha_i y_i=0 \quad ②$$

② Feasibility

$$\alpha_i\geq 0$$
$$1-y_i(\mathbf{w}^T\mathbf{x}_i+b)\leq 0 \qquad \forall i \quad ③$$

③ Complementary slackness

$$\alpha_i\left[1-y_i(\mathbf{w}^T\mathbf{x}_i+b)\right]=0 \quad \forall i \quad ④$$

- 将 ①, ② 代入 Lagrange function，有

$$
\begin{aligned}
&L(\mathbf{w},b,\boldsymbol{\alpha})\\
=\ &\frac{1}{2}\|\mathbf{w}\|^2+\sum_i^m \alpha_i(1-y_i(\mathbf{w}^T\mathbf{x}_i+b))\\
=\ &\frac{1}{2}\|\mathbf{w}\|^2+\sum_i^m \alpha_i-\sum_i^m \alpha_i y_i\Big(\sum_j^m \alpha_j y_j \mathbf{x}_j\Big)^T\mathbf{x}_i-\sum_i^m \alpha_i y_i b\\
=\ &\frac{1}{2}\|\mathbf{w}\|^2+\sum_i^m \alpha_i-\underbrace{\sum_{i,j}^m \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j}_{(\text{由①得}=\|\mathbf{w}\|^2)}-b\underbrace{\sum_i^m \alpha_i y_i}_{(\text{由②得}=0)}\\
=\ &\sum_i^m \alpha_i-\frac{1}{2}\|\mathbf{w}\|^2\\
=\ &\sum_i^m \alpha_i-\frac{1}{2}\sum_{i,j}^m \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \quad (\text{由此消去了 } \mathbf{w} \text{ 的影响})
\end{aligned}
$$

- 由此得到下述 dual problem：

$$\max_{\boldsymbol{\alpha}} \sum_i^m \alpha_i-\frac{1}{2}\sum_{i,j}^m \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$
$$\text{s.t. } \sum_i^m \alpha_i y_i=0 \qquad \forall i$$
$$\alpha_i\geq 0$$

这可以用 optimization solver 直接求解

- 最后将求出的 $\boldsymbol{\alpha}$ 代回 stationary condition，得出 the primal solution $\mathbf{w}$

$$\mathbf{w}=\sum_i^m \alpha_i y_i \mathbf{x}_i$$

2、Solution intepretation

Solution interpretation:

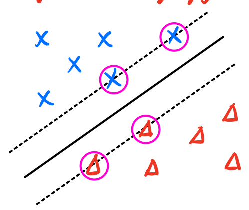
- The primal solution $\mathbf{w}$ and the dual solution $\boldsymbol{\alpha}$ should also satisfy other KKT conditions
  - Feasibility: $\alpha_i \geq 0,\ 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 0,\ \forall i$
  - Complementary slackness: $\alpha_i\big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big) = 0,\ \forall i$
- When comparing above conditions together, we have that for $\mathbf{x}_i,\ \forall i$,
  - If it satisfies $1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) < 0$, then $\alpha_i = 0$;
  - If it satisfies $1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 0$, then $\alpha_i \geq 0$.

若 $\alpha_i=0$，则 $\mathbf{x}_i$ 不会对 $\mathbf{w}_i$ 产生影响.

然而实际上，大多数 α_i 均为 0，仅有 $1-y_i(\mathbf{w}^T\mathbf{x}_i+b)=0$ 且 $\alpha_i>0$ 的 $\mathbf{x}_i$ 才能 construct the classifier.

这些 $\mathbf{x}_i$ 被称为 support vectors，位于超平面 $y_i(\mathbf{w}^T\mathbf{x}_i+b)=1$ 上

定义 $S=\{i \mid \alpha_i>0\}$ 为 support set



3、Optimization of SVM (b)

根据上述分析，对任意 support vector $\mathbf{x}_j$，$j\in S$，有

$$y_j(\mathbf{w}^T\mathbf{x}_j+b)=1, \quad \forall j\in S$$

$$\Rightarrow y_j\Big(\sum_i^m \alpha_i y_i \mathbf{x}_i^T \mathbf{x}_j + b\Big) = 1$$

由于 $y_j\cdot y_j=1$ ($y_j=\pm1$)，我们有

$$\sum_i^m \alpha_i y_i \mathbf{x}_i^T\mathbf{x}_j + b = y_j$$

因此用上所有 support vectors (也可以仅用一个)，我们可以求出 b：

$$b=\frac{1}{|S|}\sum_{j\in S}\Big(y_j-\sum_i^m\alpha_i y_i \mathbf{x}_i^T\mathbf{x}_j\Big)$$

4、Prediction using SVM

给定 optimized parameters $\{\boldsymbol{\alpha}, \mathbf{w}, b\}$，给定一个 new data $\mathbf{x}$，prediction 为

$$\mathbf{w}^T\mathbf{x}+b = \sum_i^m \alpha_i y_i \mathbf{x}_i^T\mathbf{x} + \frac{1}{|S|}\sum_{j\in S}\Big(y_j-\sum_i^m\alpha_i y_i \mathbf{x}_i^T\mathbf{x}_j\Big)$$

- 若 $\mathbf{w}^T\mathbf{x}+b>0$，则 $\mathbf{x}$ 的 predicted class 为 $+1$，otherwise -1
- 当且仅当 $y(\mathbf{w}^T\mathbf{x}+b)>0$ 时，prediction 为 correct