Logistic Regression Classification

Logistic Regression Classification

- Consider binary classification:
  - $y = 0, 1$
  - Each example represented by a feature vector $\mathbf{x}$
- Intuition: map $\mathbf{x}$ to a real number → $\mathbf{w}^\top \mathbf{x}$
  - Very positive $\mathbf{w}^\top \mathbf{x}$ means $\mathbf{x}$ is likely in the positive class $(y = 1)$
  - Very negative $\mathbf{w}^\top \mathbf{x}$ means $\mathbf{x}$ is likely in the negative class $(y = 0)$
- Probability interpretation: $\mathbf{w}^\top \mathbf{x} \rightarrow p(y|\mathbf{x})$
- Squash the range of $\mathbf{w}^\top \mathbf{x} \in (-\infty, +\infty)$ down to $[0, 1]$

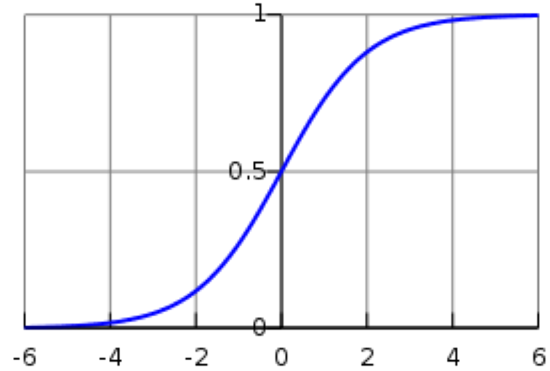
Logistic Regression Classification

- Probability interpretation: $\mathbf{w}^\top \mathbf{x} \rightarrow p(y|\mathbf{x})$

Conditional Probability: relevant in classification

$\sigma(z) = \frac{1}{1+e^{-z}}$ Logistic function / sigmoid function

$z \rightarrow +\infty, \sigma(z) \rightarrow 1; z \rightarrow -\infty, \sigma(z) \rightarrow 0$



$$p(y=1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1+exp(-\mathbf{w}^\top \mathbf{x})} = \frac{exp(\mathbf{w}^\top \mathbf{x})}{1+exp(\mathbf{w}^\top \mathbf{x})}$$

$$p(y=0|\mathbf{x}) = 1 - p(y=1|\mathbf{x}) = \frac{1}{1+exp(\mathbf{w}^\top \mathbf{x})}$$

Logistic Regression: Log Odds

- 一个事件的几率(odds)：
  - 该事件发生的概率与不发生的概率的比值，p/(1-p)
  - log odds / logit function: log[p/(1-p)]

- Log odds for logistic regression:

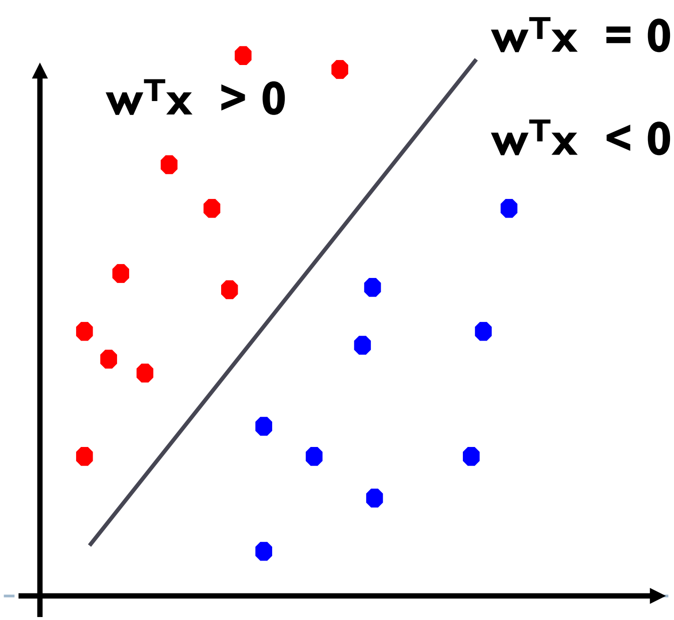
$$\log \frac{p(y=1|\mathbf{x})}{1-p(y=1|\mathbf{x})} = \mathbf{w}^\top \mathbf{x}$$

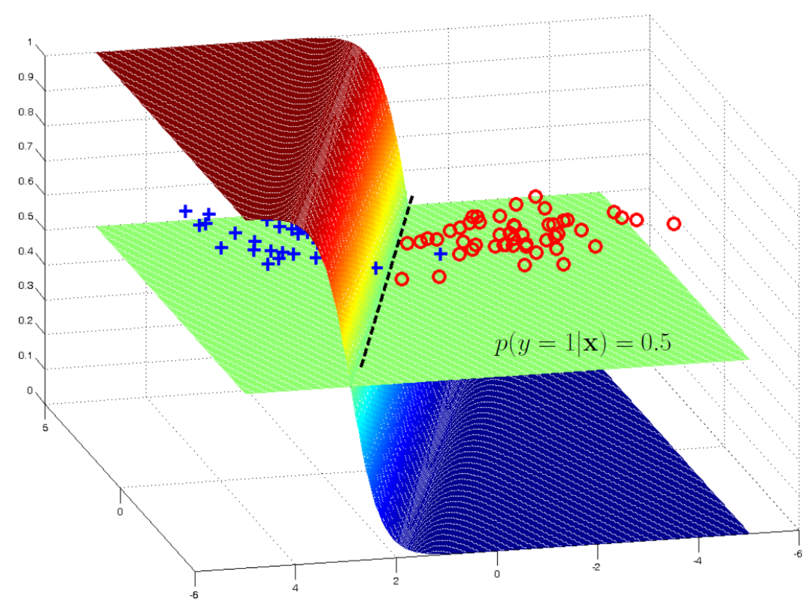
Logistic Regression: Decision Boundary

If $p(y=1|\mathbf{x}) \geq 0.5$, predict $y=1$

If $p(y=1|\mathbf{x}) < 0.5$, predict $y=0$

- Decision boundary: $p(y=1|\mathbf{x}) = 0.5 \Leftrightarrow \mathbf{w}^\top \mathbf{x} = 0$
- linear logistic model → a linear decision boundary





Likelihood under the Logistic Model

- Logistic regression: observe labels, measure their probability under the model

给定模型w，每个样本属于其真实类别的概率。

$$
\begin{aligned}
p(y_i|\mathbf{x}_i;\mathbf{w}) &= \begin{cases} \sigma(\mathbf{w}^\top \mathbf{x}_i) & \text{if} \quad y_i = 1, \\ 1-\sigma(\mathbf{w}^\top \mathbf{x}_i) & \text{if} \quad y_i = 0 \end{cases} \\
&= \sigma(\mathbf{w}^\top \mathbf{x}_i)^{y_i}(1-\sigma(\mathbf{w}^\top \mathbf{x}_i))^{1-y_i}
\end{aligned}
$$

- The conditional log-likelihood of w:

$$
\begin{aligned}
\ell(\mathbf{w}) &= \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i;\mathbf{w}) \\
&= \sum_{i=1}^{N} y_i \log \sigma(\mathbf{w}^\top \mathbf{x}_i) + (1-y_i)\log\left(1-\sigma(\mathbf{w}^\top \mathbf{x}_i)\right)
\end{aligned}
$$

Training the Logistic Model

- Training (i.e., finding the parameter w) can be done by maximizing the conditional log likelihood of training data $\{(\mathbf{x}_i, y_i)\}_{i=1:N}$

$$\max_{\mathbf{w}} \ell(\mathbf{w}) = \max_{\mathbf{w}} \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w})$$

or

$$\begin{aligned} \min_{\mathbf{w}} J(\mathbf{w}) &= \min_{\mathbf{w}} -\ell(\mathbf{w}) \\ &= \min_{\mathbf{w}} -\left[\sum_{i=1}^{N} y_i \log \sigma(\mathbf{w}^\top \mathbf{x}_i) + (1 - y_i) \log\left(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)\right)\right] \end{aligned}$$

Gradient Descent

- Want $\min_{\mathbf{w}} J(\mathbf{w})$

Repeat {

$$\mathrm{w}_j := \mathrm{w}_j - \alpha \frac{\partial}{\partial \mathrm{w}_j} J(\mathbf{w})$$

} (simultaneously update all w_j)

Homework: Derivative of the Logistic

- A useful fact

$$\frac{\partial}{\partial z}\sigma(z) = \frac{\partial}{\partial z}\frac{1}{1+e^{-z}} = \underbrace{-\left(\frac{1}{1+e^{-z}}\right)^2}_{\partial\sigma/\partial(1+e^{-z})} \times \underbrace{-e^{-z}}_{\partial(1+e^{-z})/\partial z}$$

$$= \sigma^2(z)\left(\frac{1-\sigma(z)}{\sigma(z)}\right) = \sigma(z)(1-\sigma(z)).$$

- Compute $\frac{\partial}{\partial \mathrm{w}_j} J(\mathbf{w})$

Comments on Logistic Regression

- Parametric learning model

- Linear classification

- Discriminative model: estimate conditional likelihood p(y|x) directly