Logistic Regression

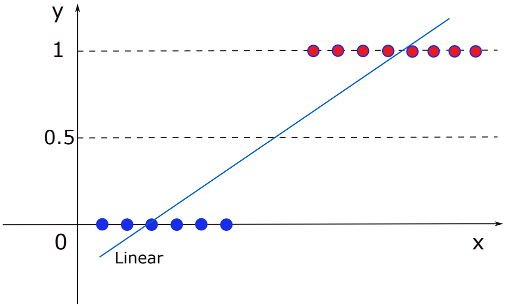
By Van Dinh Tran

February 16, 2025

Linear regression as classification?

- A probabilistic classifier's output:
$$h(x) = P(y = 1|x)$$
It outputs the probability rather than the label of the most likely class.

- Why linear regression, $y = w^{\intercal}x$, is not relevant for classification?
  - In regression, $y_i \in R$, in classification it is categorical.
  - The output of the linear regression model can be out of the [0,1] range.



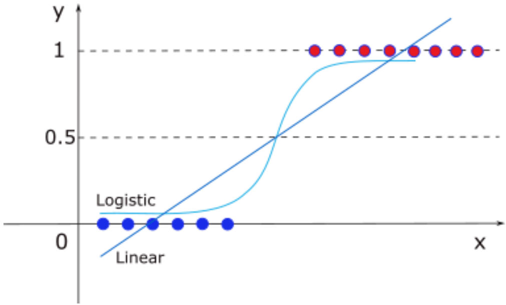
Logistic regression (LR)

- Logistic regression with the use of the Sigmoid function is better suited for classification which can output the probability.

$$\mathrm{Sigmoid}(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

- It uses linear regression as the basis

$$\mathrm{Sigmoid}(w^{\intercal}x) = \frac{1}{1+e^{-w^{\intercal}x}} = \frac{e^{w^{\intercal}x}}{1+e^{w^{\intercal}x}}$$



Logistic regression (LR)

- Hypothesis: $h_w(x) = p(x) = \frac{1}{1+e^{-w^\intercal x}}$
    - If $p(x) \geq 0.5$, $y = 1$, otherwise 0
- Odds $= \frac{p(x)}{1-p(x)} \Longrightarrow \text{Log}(\text{odds}) = w^\intercal x$

  $p(x)$ is proportional with Log(odds), i.e. $w^\intercal x$

We aim to estimate w such that $p(x)$ is as close to 1 as possible

Let $p(y=1|x) = p(x) \Longrightarrow p(y=0|x) = 1 - p(x)$

We can combine them in a compact form as follows.

$$p(y|x) = p(x)^y . (1 - p(x))^{1-y}$$

Likelihood function

- Likelihood function, $\mathcal{L}(w)$:

$$\mathcal{L}(w) = \prod_{i=1}^{N} p(y_i|x_i) = \prod_{i=1}^{N} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

- Log-likelihood, $\ell(w)$:

$$\begin{aligned} \ell(w) &= \mathrm{Log}(\prod_{i=1}^{N} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}) \\ &= \sum_{i=1}^{N} [y_i w^{\intercal} x_i - \log(1 + e^{w^{\intercal} x_i})] \end{aligned}$$

- Objective function:

$$\mathrm{Argmax}_w \ell(w)$$

- Taking derivative:

$$\frac{\partial \ell(w)}{\partial w} = \sum_{i=1}^{N} x_i (y_i - p(x_i))$$

Gradient Ascent

$$\text{Argmax}_w \ell(w)$$

Note that given $p(x_i)$, we can calculate the derivative. Thus, Gradient ascent can be adopted to solve this optimization.

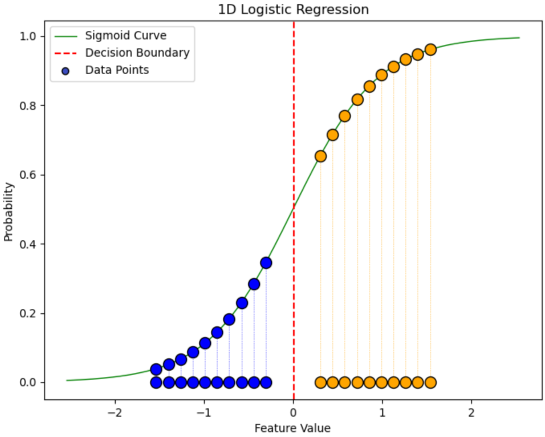
Gradient Ascent algorithm

- Pick some value for w
- Iteratively update w until convergence:

$$w \leftarrow w + \alpha * \frac{\partial \ell(w)}{\partial w}$$

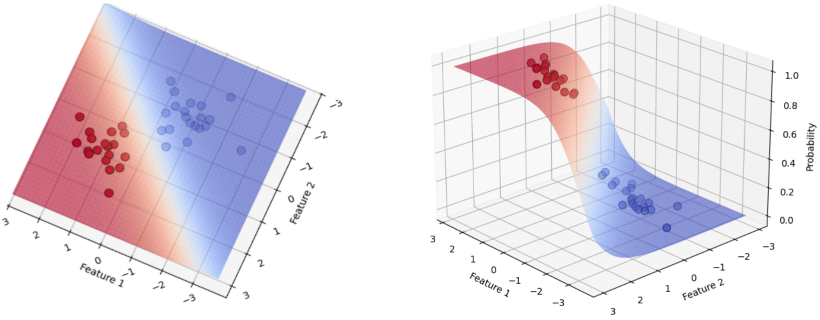
in which $\frac{\partial \ell(w)}{\partial w} = \sum_{i=1}^{N} x_i(y_i - p(x_i))$

Visualization



Logistic Regression on 1D data

Visualization



Logistic Regression on 2D data

Note that the decision curve in Logistic regression is a collection of data points x_i such that $h(x_i) = 0.5$

Pros and cons of LR

- **Advantages**
  - Simplicity and interpretability
  - Computationally efficient
  - Probabilistic output
- **Disadvantages**
  - Incapable of dealing well with non-linear relationships
  - Vulnerable to overfitting
  - Sensitive to outliers

Supplementary

$$\mathcal{L}(\mathrm{w}) = \prod_{i=1}^{N} p(y_i|x_i) = \prod_{i=1}^{N} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

Objective function: $\mathrm{Max}_{\mathrm{w}}\mathcal{L}(\mathrm{w})$

Log-likelihood:

$$\begin{aligned}
\ell(\mathrm{w}) &= \mathrm{Log}(\mathcal{L}(\mathrm{w})) \\
&= \mathrm{Log}(\prod_{i=1}^{N} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}) \\
&= \sum_{i=1}^{N}[y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))] \\
&= \sum_{i=1}^{N}[y_i(\log(p(x_i)) - \log(1 - p(x_i))) + \log(1 - p(x_i))]
\end{aligned}$$

Log-likelihood

$$
\begin{aligned}
\ell(w) &= \sum_{i=1}^{N}[y_i(\log(p(x_i)) - \log(1 - p(x_i))) + \log(1 - p(x_i))] \\
&= \sum_{i=1}^{N}[y_i \log(\frac{p(x_i)}{1 - p(x_i)}) + \log(1 - p(x_i))] \\
&= \sum_{i=1}^{N}[y_i w^\intercal x_i + \log(1 - \frac{e^{w^\intercal x_i}}{1 + e^{w^\intercal x_i}}))] \\
&= \sum_{i=1}^{N}[y_i w^\intercal x_i + \log(\frac{1}{1 + e^{w^\intercal x_i}})] \\
&= \sum_{i=1}^{N}[y_i w^\intercal x_i - \log(1 + e^{w^\intercal x_i})]
\end{aligned}
$$

Taking derivative

$$
\begin{aligned}
\frac{\partial \ell(\mathrm{w})}{\partial \mathrm{w}} &= \sum_{i=1}^{N} [y_i x_i - \frac{1}{1+e^{w^\intercal x_i}} . e^{w^\intercal x_i} . x_i] \\
&= \sum_{i=1}^{N} [y_i x_i - \frac{e^{w^\intercal x_i}}{1+e^{w^\intercal x_i}} . x_i] \\
&= \sum_{i=1}^{N} [y_i x_i - p(x_i) . x_i] \\
&= \sum_{i=1}^{N} x_i (y_i - p(x_i))
\end{aligned}
$$