Logistic Regression

Observation: \( Y \in \{0, 1\} \)
- e.g. presence or absence of disease

Feature: \( X = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \in \mathbb{R}^p \)

Regression coefficient: \( \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^p \)

For data with binary observation, the likelihood function is
\[ P(Y = y_i \mid X_i = x_i) = p(x_i, \beta)^{y_i}(1 - p(x_i, \beta))^{1 - y_i} \]

Why this likelihood?
\[
\begin{align*}
P(Y = 1 \mid X_i = x_i) &= p(x_i, \beta) \\
P(Y = 0 \mid X_i = x_i) &= 1 - p(x_i, \beta)
\end{align*}
\]

Therefore, by independence we have
\[
\prod_{i=1}^{N} P(Y = y_i \mid X_i = x_i) = \prod_{i=1}^{N} p(x_i, \beta)^{y_i}(1 - p(x_i, \beta))^{1 - y_i}
\]

Now, we need a model for \( p(x_i, \beta) \).

Option 1: \( p(x_i, \beta) = \beta_0 + \beta^T x_i \)
- This does not work because
  - \( p(x_i, \beta) \) must be probability so that \( 0 < p(x_i, \beta) < 1 \)
  - But if \( p(x_i, \beta) = \beta_0 + \beta^T x_i \), then \( p(x_i, \beta) \to \pm \infty \)
Option 2 \[ \log P(x_i, \beta) = \beta_0 + \beta^T x_i \]

This is better but still doesn't work

\[ p(x_i, \beta) > 0 \text{ because } \beta_0 + \beta^T x_i \to -\infty \Rightarrow \log p \to 0. \]

But unbounded above.

Option 3 \[ \log \frac{P(x_i, \beta)}{1 - P(x_i, \beta)} = \beta_0 + \beta^T x_i. \]

Then

\[ P(x_i, \beta) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x_i)}} \]

Three properties:

(i) \[ 0 < P(x, \beta) < 1 \]

(ii) decision boundary is at \[ \beta_0 + \beta^T x = 0 \]

(iii) rate of sharpness of decision depends on \( ||\beta|| \).

**How to estimate \( \beta \) from logit model?**

Recall

\[ P(Y = 1 \mid X_i = x_i) = P(x_i, \beta) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x_i)}} \]

So the negative log likelihood is

\[ -\log \prod_{i=1}^{N} \left\{ \frac{1}{1 + e^{-(\beta_0 + \beta^T x_i)}} \right\} \]

\[ = -\log \frac{1}{\prod_{i=1}^{N} \left\{ \frac{1}{1 + e^{-(\beta_0 + \beta^T x_i)}} \right\}} \]

\[ = -\sum_{i=1}^{N} \left\{ y_i \log \frac{p(x_i, \beta)}{1 - p(x_i, \beta)} + (1 - y_i) \log \left(1 - p(x_i, \beta) \right) \right\} \]

\[ = -\sum_{i=1}^{N} \left\{ y_i \log \frac{p(x_i, \beta)}{1 - p(x_i, \beta)} + \log \left(1 - p(x_i, \beta) \right) \right\} \]
\[= - \sum_{i=1}^{N} \left\{ y_i (\beta_0 + \beta^T x_i) + \log \left( \frac{1}{1 + e^{\beta_0 + \beta^T x_i}} \right) \right\} \]

\[= - \sum_{i=1}^{N} \left\{ y_i (\beta_0 + \beta^T x_i) - \log \left( 1 + e^{\beta_0 + \beta^T x_i} \right) \right\} \]

So we can take derivative:
\[\frac{\partial}{\partial \beta}(\cdot) = 0 \Rightarrow \text{transcendental equation} \]
(can be solved numerically).

To include sparsity in the logistic regression, we add
\[\min_{(\beta_0, \beta)} \left\{ -\sum_{i=1}^{N} \left[ y_i (\beta_0 + \beta^T x_i) - \log \left( 1 + e^{\beta_0 + \beta^T x_i} \right) \right] + \lambda \| \beta \|_1 \right\} \]

---

**Poisson Regression**

**Observation:** \(Y = 0, 1, 2, \ldots\)

The likelihood function is
\[
\Pr(Y = y_i \mid X_i = x_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}
\]

where \(\lambda_i\) is a function of \(x_i, \beta\).

(So technically should be \(\lambda_i = \lambda (x_i, \beta)\)).

Therefore, where there are \(N\) observations,
\[
\prod_{i=1}^{N} \Pr(Y = y_i \mid X_i = x_i) = \prod_{i=1}^{N} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} = \prod_{i=1}^{N} \lambda_i^{y_i} e^{-\lambda_i} \frac{1}{y_i!}
\]
What should be the model of $\lambda(x_i, \beta)$?
Again, we choose
\[
\log \left\{ \lambda(x_i, \beta) \right\} = \beta_0 + \beta^T x_i .
\]

Remark:
\[
\lambda(x_i, \beta) = \mathbb{E}(Y \mid X = x_i).
\]
So (1) can be written as
\[
\log \mathbb{E}(Y \mid X = x_i) = \beta_0 + \beta^T x_i .
\]

How to solve $\beta$ from Poisson Regression?
The negative log-likelihood is
\[
- \log P(Y = y \mid X = x) = - \log \left\{ \frac{\prod_{i=1}^{N} \lambda(x_i, \beta)^{y_i} e^{\lambda(x_i, \beta)}}{y_i !} \right\}
\]

\[
= - \sum_{i=1}^{N} \left\{ y_i \log \lambda(x_i, \beta) + \lambda(x_i, \beta) - \log(y_i !) \right\}
\]

By (1)
\[
= - \sum_{i=1}^{N} \left\{ y_i \left( \beta_0 + \beta^T x_i \right) + e^{\beta_0 + \beta^T x_i} - \log(y_i !) \right\}
\]

So $\beta$ can be found by
\[
(\hat{\beta}_0, \hat{\beta}) = \arg \min_{(\beta_0, \beta)} - \sum_{i=1}^{N} \left\{ y_i \left( \beta_0 + \beta^T x_i \right) + e^{\beta_0 + \beta^T x_i} - \log(y_i !) \right\}.
\]

If we want sparsity, then add $\|\beta\|_1$. "Note: we do not need to upper-bound $\lambda(x_i, \beta)$, and so we do not need things like $\frac{\lambda(x_i, \beta)}{1 - \lambda(x_i, \beta)}$.}
**Generalized Linear Model (GLM)**

From the above examples, we see that for data that cannot be modeled by the standard linear model, the GLM could become very useful. In general, the transformation of the conditional mean

$$g\left( \mathbb{E}[Y \mid X = x] \right) \overset{\text{def}}{=} \mu(x)$$

is called a **link function**.

**Logit**: \[ \mathbb{E}[Y \mid X = x] = \frac{\mathbb{P}(Y=1 \mid X = x)}{1 + \mathbb{P}(Y=0 \mid X = x)} \]

Note: \[ \mathbb{P}(Y=1 \mid X = x) = \mathbb{P}(x; \beta) \]

**Poisson**: \[ \mathbb{E}[Y \mid X = x] = \lambda(x; \beta) \]

More generally, we have GLM being

$$g(\mu(x)) = \beta_0 + \beta^T x$$

The choice of \( g(\cdot) \) depends on the model:

**Logit**: \( g(\mu) = \log \left( \frac{\mu}{1-\mu} \right) \)

**Poisson**: \( g(\mu) = \log \mu \)

Why is it called the **generalized linear model**? The linear model is a special case:

$$\mathbb{E}[Y \mid X = x] = \mu(x) = \beta_0 + \beta^T x$$

$$g(\mu) = \mu$$.

And if we consider a **Gaussian** likelihood:

$$\prod_{i=1}^{N} \mathbb{P}(Y = y_i \mid X_i = x_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi} \sigma^2} e^{-(y_i - \mu(x_i; \beta))^2 / 2\sigma^2}.$$
then its negative log-likelihood is

\[-\log(.) = -\left\{ \sum_{i=1}^{N} \left( \log\left( \frac{1}{\sqrt{2\pi \sigma^2}} \right) - \frac{(y_i - \mu(x_i, \beta))^2}{2\sigma^2} \right) \right\} \]

So the regression coefficients are

\[ (\hat{\beta}_0, \hat{\beta}) = \underset{\beta \in \mathbb{R}^n}{\arg \min} \sum_{i=1}^{N} (y_i - \mu(x_i, \beta))^2 \]

\[ = \underset{\beta \in \mathbb{R}^n}{\arg \min} \sum_{i=1}^{N} (y_i - (\beta_0 + \beta^T x_i))^2 \]

Support Vector Machine

\[ \beta_0 + \beta^T x = 0 \]

Idea: Find a separating hyperplane that has the maximum margin, and all the data points should be as far from the margin as possible.

But first we need a way to measure the distance between a point \( x \) and the line \( L : \beta_0 + \beta^T x = 0 \).

1. If \( x_1, x_2 \) are on \( L \), the

\[ \beta_0 + \beta^T x_1 = 0 = \beta_0 + \beta^T x_2 \]

\[ \Rightarrow \beta^T (x_1 - x_2) = 0 \]

So \( \beta^* \) def \( \frac{\beta}{\|\beta\|_2} \) is the normal.
(2) For any point \( x_0 \in L \),
\[
\beta_0 + \beta^T x_0 = 0
\]
\[\Rightarrow \beta^T x_0 = -\beta_0 .\]

(3) The distance between \( x \) and \( L \) is
\[
D = \beta^* \cdot (x - x_0), \text{ where } x_0 \text{ is a point on } L
\]
\[= \frac{\beta^T(x - x_0)}{||\beta||}
\]
\[= \frac{1}{||\beta||} (\beta^T x + \beta_0).
\]

Going back to SVM formulation, we are interested in solving the problem
\[
\begin{align*}
\max_{\beta, \beta^*} & \quad M \\
\text{s.t.} & \quad \frac{1}{||\beta||} y_i (\beta_0 + \beta^T x_i) \geq M \\
& \quad y_i > 0 \\
& \quad y_i < 0
\end{align*}
\]

Interpretation: (i) \( M \) = margin.
We want \( M \) to be as large as possible.

(ii) \( \frac{\beta_0 + \beta^T x_i}{||\beta||} \cdot y_i \) is the unsigned distance.
If \( \beta_0 + \beta^T x_i > 0 \), then ideally \( y_i > 0 \).
If \( \beta_0 + \beta^T x_i < 0 \), then ideally \( y_i < 0 \).
So if \( y_i (\beta_0 + \beta^T x_i) < 0 \), then there is misclassification.
Modification 1

\[ y_i (\beta_0 + \beta^T x_i) \geq M \|\beta\| \]  \hspace{1cm} (3)

can be simplified as

\[ y_i (\beta_0 + \beta^T x_i) \geq 1, \text{ because if } (\beta_0, \beta) \text{ satisfies (3), } \left( \frac{\beta_0}{M}, \frac{\beta}{M} \right) \text{ will satisfy (4). So we can solve the problem:} \]

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|\beta\|^2 \\
\text{subject to} & \quad y_i (\beta_0 + \beta^T x_i) \geq 1, \quad i = 1, 2, \ldots, N
\end{align*}
\]

Modification 2

We want to tolerate some misclassification. So we need to relax (4) by

\[ y_i (\beta_0 + \beta^T x_i) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N \]

and

\[ \xi_i \geq 0, \quad \sum_{i=1}^{N} \xi_i \leq \text{constant}. \]

So the SVM problem becomes

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|\beta\|^2 \\
\text{subject to} & \quad y_i (\beta_0 + \beta^T x_i) \geq 1 - \xi_i \\
& \quad \xi_i \geq 0, \quad \sum_{i=1}^{N} \xi_i \leq \text{constant}
\end{align*}
\]

This can be re-written as

\[
\begin{align*}
\text{maximize} & \quad \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} \xi_i \\
\text{subject to} & \quad y_i (\beta_0 + \beta^T x_i) \geq 1 - \xi_i \\
& \quad \xi_i \geq 0
\end{align*}
\]  \hspace{1cm} (P_i)
Let $c = \frac{1}{N}$, then the objective of $(P_1)$ becomes

$$\min_{(\beta, \beta_0)} \frac{1}{2} \| \beta \|^2 + \sum_{i=1}^{N} \xi_i$$

Note that by the constraints of $(P_1)$, we have

$$\xi_i \geq 1 - y_i (\beta_0 + \beta^T x_i)$$

and $\xi_i \geq 0$.

So

$$\sum_{i=1}^{N} \xi_i \geq \sum_{i=1}^{N} \left[ 1 - y_i (\beta_0 + \beta^T x_i) \right]_+,$$

where $[\cdot]_+$ denotes the positive part of the argument.

Therefore, $(P_1)$ can be solved by minimizing its lower bound:

$$\min_{(\beta, \beta_0)} \frac{1}{2} \| \beta \|^2 + \sum_{i=1}^{N} \left[ 1 - y_i (\beta_0 + \beta^T x_i) \right]_+.$$

And by absorbing $\frac{1}{N}$ into the equation, we have

$$\min_{(\beta, \beta_0)} \frac{1}{N} \sum_{i=1}^{N} \left[ 1 - y_i (\beta_0 + \beta^T x_i) \right]_+ + \frac{1}{2} \| \beta \|^2$$

If we want to enforce sparsity, then we can do

$$\min_{(\beta, \beta_0)} \frac{1}{N} \sum_{i=1}^{N} \left[ 1 - y_i (\beta_0 + \beta^T x_i) \right]_+ + \lambda \| \beta \|_1$$

$(P_2)$. 