Logistic Regression

Observation : $Y \in \{0, 1\}$

e.g. presence ~~of~~ or absence of disease

Feature : $\underline{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \in \mathbb{R}^p$

Regression coefficient: $\underline{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^p$.

For data with binary observation, the likelihood function is

$$\mathbb{P}(Y = y_i \mid \underline{X}_i = \underline{x}_i) = p(\underline{x}_i, \underline{\beta})^{y_i} (1 - p(\underline{x}_i, \underline{\beta}))^{1-y_i}$$

Why this likelihood?

$$\mathbb{P}(Y = 1 \mid \underline{X}_i = \underline{x}_i) = p(\underline{x}_i, \underline{\beta})$$

$$\mathbb{P}(Y = 0 \mid \underline{X}_i = \underline{x}_i) = 1 - p(\underline{x}_i, \underline{\beta})$$

Some function linking $\underline{x}_i$, $\underline{\beta}$ and the "probability"

need to define this.

Therefore, by independence we have

$$\prod_{i=1}^{N} \mathbb{P}(Y = y_i \mid \underline{X}_i = \underline{x}_i) = \prod_{i=1}^{N} p(\underline{x}_i, \underline{\beta})^{y_i} (1 - p(\underline{x}_i, \underline{\beta}))^{1-y_i}$$

Now, we need a model for $p(\underline{x}_i, \underline{\beta})$.

<u>Option 1</u> : $p(\underline{x}_i, \underline{\beta}) = \beta_0 + \underline{\beta}^T \underline{x}_i$.

This does not work because

$p(\underline{x}_i, \underline{\beta})$ must be probability so that

$0 < p(\underline{x}_i, \underline{\beta}) < 1$. But if $p(\underline{x}_i, \underline{\beta}) = \beta_0 + \underline{\beta}^T \underline{x}_i$, then

$p(\underline{x}_i, \underline{\beta}) \longrightarrow \pm \infty$.

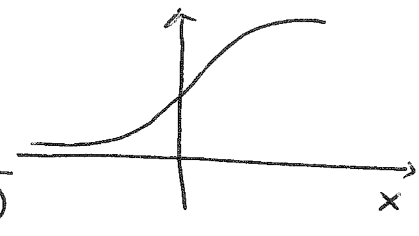
<u>option 2</u> $\log p(\underline{x}_i, \underline{\beta}) = \beta_0 + \underline{\beta}^T \underline{x}_i$

This is better but still doesn't work

$p(\underline{x}_i, \underline{\beta}) > 0$ because $\beta_0 + \underline{\beta}^T\underline{x}_i \to -\infty \Rightarrow \log p \to 0$.

But unbounded above.

<u>option 3</u> $\log \dfrac{p(\underline{x}_i, \underline{\beta})}{1 - p(\underline{x}_i, \underline{\beta})} = \beta_0 + \underline{\beta}^T \underline{x}_i$.



Then
$$p(\underline{x}_i, \underline{\beta}) = \frac{1}{1 + e^{-(\beta_0 + \underline{\beta}^T \underline{x}_i)}}$$

Three properties:

(i) $0 < p(\underline{x}, \underline{\beta}) < 1$

(ii) decision boundary is at $\beta_0 + \underline{\beta}^T \underline{x} = 0$

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

How to estimate $\underline{\beta}$ from Logistic model?

Recall $\mathbb{P}(Y = 1 \mid \underline{X}_i = \underline{x}_i) = p(\underline{x}_i, \underline{\beta}) = \dfrac{1}{1 + e^{-(\beta_0 + \underline{\beta}^T \underline{x}_i)}}$

$$= \frac{e^{\beta_0 + \underline{\beta}^T \underline{x}_i}}{1 + e^{\beta_0 + \underline{\beta}^T \underline{x}_i}}$$

So the negative-log likelihood is

$$-\log \mathbb{P}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x})$$
$$= -\log \left\{ \prod_{i=1}^{N} p(\underline{x}_i, \underline{\beta})^{y_i} \left(1 - p(\underline{x}_i, \underline{\beta})\right)^{1 - y_i} \right\}$$
$$= -\sum_{i=1}^{N} \left\{ y_i \log p(\underline{x}_i, \underline{\beta}) + (1 - y_i) \log\left(1 - p(\underline{x}_i, \underline{\beta})\right) \right\}$$
$$= -\sum_{i=1}^{N} \left\{ y_i \log \frac{p(\underline{x}_i, \underline{\beta})}{1 - p(\underline{x}_i, \underline{\beta})} + \log\left(1 - p(\underline{x}_i, \underline{\beta})\right) \right\}$$

$$= -\sum_{i=1}^{N}\left\{ y_i(\beta_0+\underline{\beta}^T\underline{x}_i) + \log\left(\frac{1}{1+e^{\beta_0+\underline{\beta}^T\underline{x}_i}}\right)\right\}$$

$$= -\sum_{i=1}^{N}\left\{ y_i(\beta_0+\underline{\beta}^T\underline{x}_i) - \log\left(1+e^{\beta_0+\underline{\beta}^T\underline{x}_i}\right)\right\}$$

So we can take derivative:

$$\frac{\partial}{\partial \underline{\beta}}(\cdot)=0 \implies \text{transcendental equation}$$
(can be solved numerically).

To include sparsity in the logistic regression, we add

$$\min_{(\beta_0,\underline{\beta})}\left\{-\sum_{i=1}^{N}\left\{y_i(\beta_0+\underline{\beta}^T\underline{x}_i) - \log\left(1+e^{\beta_0+\underline{\beta}^T\underline{x}_i}\right)\right\} + \lambda\|\beta\|_1\right\}.$$

Poisson Regression

Observation: $Y = 0, 1, 2, \ldots$

The likelihood function is

$$\mathbb{P}(Y=y_i \mid \underline{X}_i=\underline{x}_i) = \frac{\lambda_i^{y_i} e^{\lambda_i}}{y_i!},$$

where λ_i is a function of $\underline{x}_i, \underline{\beta}$.

(So technically should be ~~λ~~
$\lambda_i = \lambda(\underline{x}_i, \underline{\beta})$.)

Therefore, where there are N observations.

$$\prod_{i=1}^{N}\mathbb{P}(Y=y_i \mid \underline{X}_i=\underline{x}_i) = \prod_{i=1}^{N}\frac{\lambda(\underline{x}_i,\underline{\beta})^{y_i}\, e^{\lambda(\underline{x}_i,\underline{\beta})}}{y_i!}$$

What should be the model of $\lambda(\underline{x}_i, \beta)$?

Again, we choose

$$\log\{\lambda(\underline{x}_i, \beta)\} = \beta_0 + \beta^T \underline{x}. \qquad \text{——— (1)}$$

Note: we do not need to upperbound $\lambda(\underline{x}_i, \beta)$, and so we do not need things like $\frac{\lambda(\underline{x}_i, \beta)}{1 - \lambda(\underline{x}_i, \beta)}$.

Remark:

$$\lambda(\underline{x}_i, \beta) = \mathbb{E}[Y \mid X = x_i].$$

So (1) can be written as

$$\log \mathbb{E}[Y \mid X = x_i] = \beta_0 + \beta^T \underline{x}.$$

How to solve β from Poisson Regression?

The negative log-likelihood is

$$\begin{aligned}
&- \log \mathbb{P}(\underline{Y} = \underline{y} \mid \underline{X} = \underline{x}) \\
&= -\log\left\{ \prod_{i=1}^{N} \frac{\lambda(\underline{x}_i, \beta)^{y_i} e^{\lambda(\underline{x}_i, \beta)}}{y_i !} \right\} \\
&= - \sum_{i=1}^{N} \left\{ y_i \log \lambda(\underline{x}_i, \beta) + \lambda(\underline{x}_i, \beta) - \log(y_i !) \right\} \\
&\overset{\text{by (1)}}{=} - \sum_{i=1}^{N} \left\{ y_i (\beta_0 + \beta^T \underline{x}_i) + e^{\beta_0 + \beta^T \underline{x}_i} - \log(y_i !) \right\}
\end{aligned}$$

So β can be found by

$$(\hat{\beta}_0, \hat{\beta}) = \underset{(\beta_0, \beta)}{\operatorname{argmin}} - \sum_{i=1}^{N} \left\{ y_i (\beta_0 + \beta^T \underline{x}_i) + e^{\beta_0 + \beta^T \underline{x}_i} \right\}.$$

If we want sparsity, then add $\lambda \|\beta\|_1$.

Generalized Linear Model (GLM)

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

$$g(\mathbb{E}[Y \mid \underline{X} = \underline{x}]) \quad \stackrel{\text{def}}{=} \mu(\underline{x})$$

is called a link function.

$$\text{Logistic}: \quad \mathbb{E}[Y \mid \underline{X} = \underline{x}] = 1\, \mathbb{P}(Y=1 \mid \underline{X}=\underline{x}) + 0\, \mathbb{P}(Y=0 \mid \underline{X}=\underline{x})$$
$$= \mathbb{P}(Y=1 \mid \underline{X} = \underline{x}) = p(\underline{x}; \underline{\beta})$$

$$\text{Poisson}: \quad \mathbb{E}[Y \mid \underline{X} = \underline{x}] = \lambda(\underline{x}, \underline{\beta}).$$

More generally, we have GLM being

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

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

$$\text{Logistic}: \quad g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

$$\text{Poisson}: \quad g(\mu) = \log \mu$$

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

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

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

And if we consider a Gaussian likelihood:

$$\prod_{i=1}^{N} \mathbb{P}(\underline{Y} = y_i \mid \underline{X}_i = \underline{x}_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu(\underline{x}_i, \underline{\beta}))^2}{2\sigma^2}},$$

then its negative log-likelihood is

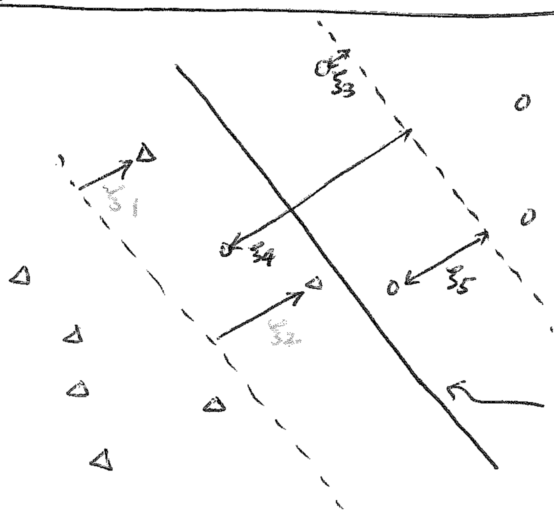
$$-\log(\cdot) = -\left\{ \sum_{i=1}^{N} \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - \mu(\underline{x}_i, \beta))^2}{2\sigma^2} \right) \right\}$$

So the regression coefficients are

$$(\hat{\beta}_0, \hat{\beta}) = \underset{(\beta_0, \beta)}{\arg\min} \sum_{i=1}^{N} (y_i - \mu(x_i, \beta))^2$$

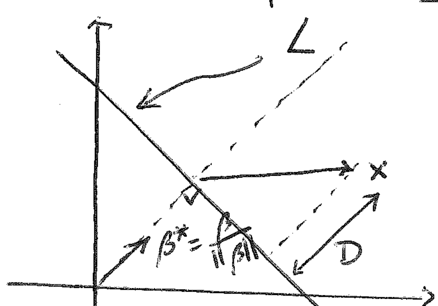
$$= \underset{(\beta_0, \beta)}{\arg\min} \sum_{i=1}^{N} (y_i - (\beta_0 + \beta^T \underline{x}_i))^2 .$$

Support Vector Machine



__Idea__: Find a separating hyperplane that has the ~~minimum~~ 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 $\underline{x}$ and the line $L: \beta_0 + \beta^T \underline{x} = 0$.



(1) if $\underline{x}_1, \underline{x}_2$ are on L, then

$$\beta_0 + \beta^T \underline{x}_1 = 0 = \beta_0 + \beta^T \underline{x}_2$$

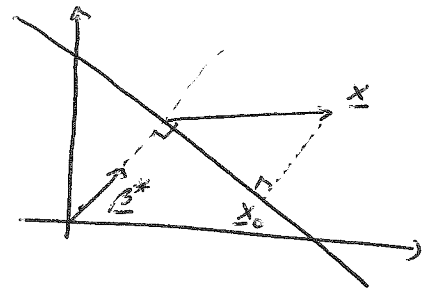
$$\Longrightarrow \beta^T (\underline{x}_1 - \underline{x}_2) = 0$$

So $\beta^* \overset{def}{=} \frac{\beta}{\|\beta\|_2}$ is the normal

(2) For any point $x_0 \in L$,

$$\beta_0 + \underline{\beta}^T \underline{x}_0 = 0$$

$$\Longrightarrow \quad \underline{\beta}^T \underline{x}_0 = -\beta_0 .$$



(3) The distance between $\underline{x}$ and L is

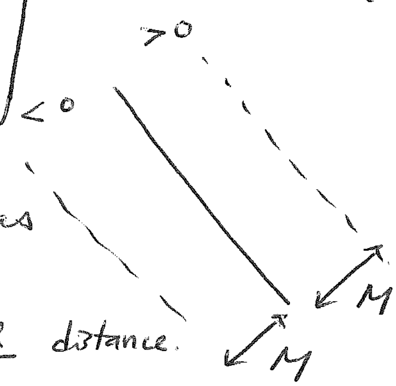
$$D = \beta^{*T}(\underline{x} - \underline{x}_0), \quad \text{where } \underline{x}_0 \text{ is a point on } L$$

$$= \frac{\underline{\beta}^T(\underline{x} - \underline{x}_0)}{\|\underline{\beta}\|}$$

$$= \frac{1}{\|\underline{\beta}\|}(\underline{\beta}^T \underline{x} + \beta_0) .$$ ← this is a signed distance

Going back to SVM formulation, we are interested in solving the problem

$$\boxed{\begin{aligned} &\max_{\beta_0, \underline{\beta}} \; M \\ &\text{s.t. } \frac{1}{\|\underline{\beta}\|} y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) \geq M \end{aligned}} \qquad (2)$$



Interpretation: (i) M = margin

we want M to be as large as possible

(ii) $\dfrac{\beta_0 + \underline{\beta}^T \underline{x}_i}{\|\beta\|} \cdot y_i$ is the <u>unsigned</u> distance.

if $\beta_0 + \underline{\beta}^T \underline{x}_i > 0$, then ideally $y_i > 0$

if $\beta_0 + \underline{\beta}^T \underline{x}_i < 0$, then ideally $y_i < 0$.

so if $y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) < 0$, then there is misclassification.

Modification 1

$$y_i(\beta_0 + \beta^T \underline{x}_i) \geq M\|\beta\| \qquad (3)$$

can be simplified as

$$y_i(\beta_0 + \beta^T \underline{x}_i) \geq 1 \qquad (4)$$

because if (β_0, β) satisfies (3), $(\frac{1}{M}\beta_0, \frac{1}{M}\beta)$ will satisfy (4). So we can solve this problem:

$$\boxed{\begin{aligned} &\min_{(\beta, \beta_0)} \ \tfrac{1}{2}\|\beta\|^2 \\ &\text{s.t.} \quad y_i(\beta_0 + \beta^T \underline{x}_i) \geq 1, \qquad i = 1, 2, \ldots, N \end{aligned}}$$

Modification 2

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

$$y_i(\beta_0 + \beta^T \underline{x}_i) \geq 1 - \xi_i, \qquad i = 1, 2, \ldots, N$$

and $\xi_i \geq 0$, $\sum_{i=1}^{N} \xi_i \leq$ constant.

So the SVM problem becomes

$$\boxed{\begin{aligned} &\min_{\{\xi_i\}, (\beta, \beta_0)} \ \tfrac{1}{2}\|\beta\|^2 \\ &\text{s.t.} \quad y_i(\beta_0 + \beta^T \underline{x}_i) \geq 1 - \xi_i \\ &\qquad \xi_i \geq 0, \quad \sum_{i=1}^{N} \xi_i \leq \text{constant} \end{aligned}}$$

This can be re-written as

$$\boxed{\begin{aligned} &\min_{(\beta, \beta_0), \{\xi_i\}} \ \tfrac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i \\ &\text{s.t.} \quad y_i(\beta_0 + \beta^T \underline{x}_i) \geq 1 - \xi_i \\ &\qquad \xi_i \geq 0 \end{aligned}} \qquad (P_1)$$

Let $C = \frac{1}{\lambda}$, then the objective of (P_1) becomes

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

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~~ returns the positive part of the argument.

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

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

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

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

If we wanto to enforce sparsity, then we can do

$$\boxed{\min_{(\beta, \beta_0)} \frac{1}{N}\sum_{i=1}^{N} \left[1 - y_i(\beta_0 + \beta^T x_i)\right]_+ + \lambda\|\beta\|_1} \quad (P_2).$$