Stats 503 Notes

1.1 Bayes Classifier

1.1.1 Expected loss

Law of total expectation

$$\mathbb{E}_Y(Y) = \mathbb{E}_X[\mathbb{E}_{Y|X}(Y|X)]$$

Iterated expectations

$$\mathbb{E}(Y|X_1) = \mathbb{E}_{X_2}[\mathbb{E}_{Y|X_1}(Y|X_2)|X_1]$$

$$\begin{aligned}\text{Expected Loss} &= \mathbb{E}_{XY}[L(f(X),Y)] \\ &= \sum_y \sum_x L(f(x),y)\mathbb{P}(x,y) \\ &= \int_y \int_x L(f(x),y)\mathbb{P}(x,y)\mathrm{d}x\mathrm{d}y\end{aligned} \tag{1.1.1}$$

1.1.2 Generative model

$$\mathbb{P}(y|x) = \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)} = \frac{\mathbb{P}(x|y)\mathbb{P}(y)}{\mathbb{P}(x)} \qquad \text{posterior} = \frac{joint}{evidence} = \frac{likelihood * prior}{evidence}$$

Given observable data $\vec{x}$, the goal is to minimize expected loss $\min_f \mathbb{E}_{XY}[L(f(X),Y)]$.

$$\begin{aligned}\text{Total Risk } = \text{Expected Loss} &= \mathbb{E}_{XY}[L(f(X),Y)] \\ &= \sum_y \sum_x L(f(x),y)\mathbb{P}(x,y) = \int_y \int_x L(f(x),y)\mathbb{P}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \sum_y \sum_x L(f(x),y)\mathbb{P}(x|y)\mathbb{P}(y) = \int_y \int_x L(f(x),y)\mathbb{P}(x|y)\mathbb{P}(y)\mathrm{d}x\mathrm{d}y\end{aligned} \tag{1.1.2}$$

1. Want to model joint $\mathbb{P}(x,y)$.
2. Prior $\mathbb{P}(y)$ known.
3. Assume $\mathbb{P}(x|y)$ to be a parametric model such as Normal.

4. Class-conditional $\mathbb{P}(x|y)$ are estimated from training data.
5. Compute joint $\mathbb{P}(x, y)$ from prior and class-conditional.
6. Compute or posterior $\mathbb{P}(x|y)$ using Bayes rule.
7. Identify optimal discriminant $f(\vec{x})$ by comparing posterior probabilities.

1.1.3 Discriminant model

Given observable data $\vec{x}$, the goal is to minimize expected loss $\min_f \mathbb{E}_{XY}[L(f(X), Y)]$.

$$\begin{aligned}
\text{Total Risk } &= \text{Expected Loss} = \mathbb{E}_{XY}[L(f(X), Y)] \\
&= \sum_y \sum_x L(f(x), y)\mathbb{P}(x, y) = \int_y \int_x L(f(x), y)\mathbb{P}(x, y)\mathrm{d}x\mathrm{d}y \\
&= \sum_y \sum_x L(f(x), y)\mathbb{P}(y|x)\mathbb{P}(x) = \int_y \int_x L(f(x), y)\mathbb{P}(y|x)\mathbb{P}(x)\mathrm{d}x\mathrm{d}y
\end{aligned} \tag{1.1.3}$$

1. Want to model directly posterior $\mathbb{P}(y|x)$.
2. Prior $\mathbb{P}(y)$ unknown.
3. Class-conditional $\mathbb{P}(x|y)$ unknown.
4. Model posterior $p = \mathbb{P}(y|x)$, or $logit(p) = \vec{x}^T\vec{\beta}$ from observed $\vec{x}$
5. Estimate posterior $\mathbb{P}(y|x)$ directly from data.
6. Identify optimal discriminant $f(\vec{x})$ using estimated posterior probabilities.

   Given observable data $\vec{x}$, the goal is to minimize expected loss

$$\begin{aligned}
\min_f \mathbb{E}[L(Y, f(X))] &= \min_f \int_X \int_Y L(y, f(x))\mathbb{P}(y|x)\mathbb{P}(x)\mathrm{d}y \\
&= \min_f \int_x \int_y L(y, f(x))\mathbb{P}(y|x)\mathbb{P}(x)\mathrm{d}x\mathrm{d}y \\
&= \min_f \int_x \mathbb{P}(x) \int_y L(y, f(x))\mathbb{P}(y|x)\mathrm{d}y\mathrm{d}x \\
&= \min_f \int_x \mathbb{P}(x)\mathbb{E}_{Y|X}[L(Y, f(X))|X]\mathrm{d}x \\
&= \min_f \ \mathbb{E}_X[\mathbb{E}_{Y|X}[L(Y, f(X))|X]]
\end{aligned}$$

$$\min_f \mathbb{E}[L(Y, f(X))] = \min_f \ \mathbb{E}_X[\mathbb{E}_{Y|X}[L(Y, f(X))|X]] \tag{1.1.4}$$

$$\min_f \mathbb{E}[L(Y, f(X))] = \min_f \ \mathbb{E}_X[\mathbb{E}_{Y|X}[L(Y, f(X))|X]] \tag{1.1.5}$$

1.2 Logistic Loss Function

Model label Y

- The n bi $y_1, ..., y_n$ where $y_i \in \{0, 1\}$
- The p explanatory variables for ith row are $x_{i1}, ..., x_{in}$
- Consider that $Y_i, ..., Y_n$ are independent Bernoulli random variables with parameters $p_1, ..., p_n$

The model is

$$Y_i \sim \text{Ber}(p_i),$$

where p_i is the parameter

$$\mathbb{P}(Y_i = 1) = p_i$$
$$\mathbb{P}(Y_i = 0) = 1 - p_i$$

The parameter to be estimated is p_i
The observed data is y_i,
Goal: estimate p_i from observed data y_i using Maximum Likelihood:

$$\begin{aligned} L(Y_i = y_1, ..., Y_n = y_n; p_1, ..., p_n) &= \prod_i^n \mathbb{P}(Y_i = y_i; p_i) \\ L(y^n; \vec{p}) &= \prod_i^n p_i^{y_i}(1 - p_i)^{1-y_i} \end{aligned}$$

Model parameter p

The model between x_i and p_i is linear through $\vec{\beta}$:

$$logit(p_i) = log\frac{p_i}{1 - p_i} = x_i^T\vec{\beta}$$

$$\begin{aligned} \mathbb{P}(Y_i = 1|x_i, \vec{\beta}) = p_i &= \frac{e^{x_i^T\vec{\beta}}}{1 + e^{x_i^T\vec{\beta}}} \\ \mathbb{P}(Y_i = 0|x_i, \vec{\beta}) = 1 - p_i &= \frac{1}{1 + e^{x_i^T\vec{\beta}}} \end{aligned}$$

That is, given observed x_i, we can rewrite p_i in terms of $\vec{\beta}$
New Goal: estimate $\vec{\beta}$ from observed data x_i

The Likelihood function went from

$$L(y^n; \vec{p}) = \prod_{i=1}^n p_i^{y_i}(1 - p_i)^{1-y_i}$$

to

$$L(y^n;\vec{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{x_i^T\vec{\beta}}}{1+e^{x_i^T\vec{\beta}}}\right)^{y_i} \left(\frac{1}{1+e^{x_i^T\vec{\beta}}}\right)^{1-y_i}$$

The Log Likelihood function is

$$\ell(y^n;\vec{\beta}) = \sum_{i}^{n} \left[y_i log\left(\frac{e^{x_i^T\vec{\beta}}}{1+e^{x_i^T\vec{\beta}}}\right) + (1-y_i) log\left(\frac{1}{1+e^{x_i^T\vec{\beta}}}\right)\right]$$

Maximization of Log Likelihood is

$$\max_{\vec{\beta}} \ell(y^n;\vec{\beta}) = \max_{\vec{\beta}} \sum_{i}^{n} \left[y_i log\left(\frac{e^{x_i^T\vec{\beta}}}{1+e^{x_i^T\vec{\beta}}}\right) + (1-y_i) log\left(\frac{1}{1+e^{x_i^T\vec{\beta}}}\right)\right] \tag{1.2.1}$$

Simplify further, we get

$$\begin{aligned}
\ell(y^n;\vec{\beta}) &= \sum_{i}^{n} \left[y_i log\left(\frac{e^{x_i^T\vec{\beta}}}{1+e^{x_i^T\vec{\beta}}}\right) + (1-y_i) log\left(\frac{1}{1+e^{x_i^T\vec{\beta}}}\right)\right] \\
&= \sum_{i}^{n} \left[y_i (x_i^T\vec{\beta} - log(1+e^{x_i^T\vec{\beta}})) - (1-y_i) log(1+e^{x_i^T\vec{\beta}})\right] \\
&= \sum_{i}^{n} \left[y_i x_i^T\vec{\beta} - \cancel{y_i log(1+e^{x_i^T\vec{\beta}})} - log(1+e^{x_i^T\vec{\beta}}) + \cancel{y_i log(1+e^{x_i^T\vec{\beta}})}\right] \\
&= \sum_{i}^{n} \left[y_i x_i^T\vec{\beta} - log(1+e^{x_i^T\vec{\beta}})\right]
\end{aligned}$$

Maximization of Log Likelihood becomes

$$\max_{\vec{\beta}} \ell(y^n;\vec{\beta}) = \max_{\vec{\beta}} \sum_{i}^{n} \left[y_i x_i^T\vec{\beta} - log(1+e^{x_i^T\vec{\beta}})\right] \tag{1.2.2}$$

Then, we proceeds to find $\hat{\vec{\beta}}_{MLE}$ through Newton's method and IRLS.

Sigmoid function

[**Sigmoid function / logistic function**]

$$\begin{aligned}
s(\gamma) &= \frac{1}{1+e^{-\gamma}} = \frac{e^{\gamma}}{1+e^{\gamma}} \\
s(-\gamma) &= \frac{e^{-\gamma}}{1+e^{-\gamma}} = \frac{1}{1+e^{\gamma}}
\end{aligned}$$

In terms of sigmoid function, the probabilities of Y_i are

$$\begin{aligned}
\mathbb{P}(Y_i=1|x_i,\vec{\beta}) = p_i \quad &= \frac{1}{1+e^{-x_i^T\vec{\beta}}} = s(x_i^T\vec{\beta}) \\
\mathbb{P}(Y_i=0|x_i,\vec{\beta}) = 1-p_i \quad &= \frac{1}{1+e^{x_i^T\vec{\beta}}} = s(-x_i^T\vec{\beta})
\end{aligned}$$

In terms of sigmoid function, the log likelihood function is

$$\ell(y^n;\vec{\beta}) = \sum_i^n \left[y_i log\Big(s(x_i^T\vec{\beta})\Big) + (1-y_i) log\Big(s(-x_i^T\vec{\beta})\Big)\right]$$

In terms of sigmoid function, likelihood maximization

$$\max_{\vec{\beta}} \ell(y^n;\vec{\beta}) = \max_{\vec{\beta}} \sum_i^n \left[y_i log\Big(s(x_i^T\vec{\beta})\Big) + (1-y_i) log\Big(s(-x_i^T\vec{\beta})\Big)\right] \tag{1.2.3}$$

Equivalence between Likelihood Maximization and Logistic Loss Minimization

Logistic loss function is

$$\begin{aligned} L(y, f(x)) &= log(1+e^{-yf(x)}) \quad \text{where } y \in \{-1, 1\} \\ &= log(s^{-1}(-yf(x))) \end{aligned}$$

Minimize Expected Loss

$$\min_{\vec{\beta}} \mathbb{E}[L(Y, X^T\vec{\beta})]$$

Since the joint probability density $p(\vec{x}, y)$ is unknown, we minimize the empirical risk

$$\min_{\vec{\beta}} \frac{1}{n}\Big(\sum_i^n L(y_i, x_i^T\vec{\beta})\Big)$$

Logistic Loss Minimization
The basic idea is that

$$\min_{\vec{\beta}} log\Big(s^{-1}(yf(x))\Big) = \max_{\vec{\beta}} log\Big(s(yf(x))\Big)$$

$$\begin{aligned} \min_{\vec{\beta}} \sum_i^n L(y_i, x_i^T\vec{\beta}) &= \min_{\vec{\beta}} \sum_i^n log(1+e^{-y_i x_i^T\vec{\beta}}) \\ &= \min_{\vec{\beta}} \sum_i^n log\Big(\frac{1}{\frac{1}{1+e^{-y_i x_i^T\vec{\beta}}}}\Big) = \min_{\vec{\beta}} \sum_i^n log\Big(s^{-1}(y_i x_i^T\vec{\beta})\Big) \\ &= \max_{\vec{\beta}} \sum_i^n log(\frac{1}{1+e^{-y_i x_i^T\vec{\beta}}}) = \max_{\vec{\beta}} \sum_i^n log\Big(s(y_i x_i^T\vec{\beta})\Big) \\ &= \max_{\vec{\beta}} \left[\sum_{y_i=1}^n log\Big(s(x_i^T\vec{\beta})\Big) + \sum_{y_i=-1}^n log\Big(s(-x_i^T\vec{\beta})\Big)\right] \\ &= \max_{\vec{\beta}} \left[\sum_{y_i=1}^n log\Big(\frac{1}{1+e^{-y_i x_i^T\vec{\beta}}}\Big) + \sum_{y_i=-1}^n log\Big(\frac{1}{1+e^{y_i x_i^T\vec{\beta}}})\Big)\right] \end{aligned}$$

So we get,

$$\min_{\vec{\beta}} \sum_{i}^{n} L(y_i, x_i^T\vec{\beta}) = \max_{\vec{\beta}} \Bigg[\sum_{i:y_i=1}^{n} log\Big(\frac{1}{1+e^{-x_i^T\vec{\beta}}}\Big) + \sum_{i:y_i=-1}^{n} log\Big(\frac{1}{1+e^{x_i^T\vec{\beta}}})\Big)\Bigg] \tag{1.2.4}$$

Recall from Log Likelihood Maximization where $y_i \in \{0,1\}$

$$\begin{aligned}
\max_{\vec{\beta}} \ell(y^n;\vec{\beta}) &= \max_{\vec{\beta}} \sum_{i}^{n} \Bigg[y_i log\Big(\frac{e^{x_i^T\vec{\beta}}}{1+e^{x_i^T\vec{\beta}}}\Big) + (1-y_i)log\Big(\frac{1}{1+e^{x_i^T\vec{\beta}}}\Big)\Bigg] \\
&= \max_{\vec{\beta}} \sum_{i}^{n} \Bigg[y_i log\Big(\frac{1}{1+e^{-x_i^T\vec{\beta}}}\Big) + (1-y_i)log\Big(\frac{1}{1+e^{x_i^T\vec{\beta}}}\Big)\Bigg] \\
&= \max_{\vec{\beta}} \Bigg[\sum_{y_i=1}^{n} y_i log\Big(\frac{1}{1+e^{-x_i^T\vec{\beta}}}\Big) + \sum_{y_i=0}^{n} (1-y_i)log\Big(\frac{1}{1+e^{x_i^T\vec{\beta}}}\Big)\Bigg] \\
&= \max_{\vec{\beta}} \Bigg[\sum_{y_i=1}^{n} log\Big(\frac{1}{1+e^{-x_i^T\vec{\beta}}}\Big) + \sum_{y_i=0}^{n} log\Big(\frac{1}{1+e^{x_i^T\vec{\beta}}}\Big)\Bigg]
\end{aligned}$$

So we have,

$$\max_{\vec{\beta}} \ell(y^n;\vec{\beta}) = \max_{\vec{\beta}} \Bigg[\sum_{i:y_i=1}^{n} log\Big(\frac{1}{1+e^{-x_i^T\vec{\beta}}}\Big) + \sum_{i:y_i=0}^{n} log\Big(\frac{1}{1+e^{x_i^T\vec{\beta}}}\Big)\Bigg] \tag{1.2.5}$$

The binary responses are observed, so the size of each class remains the same regardless of how they are labeled, so Equation 1.2.4 and Equation 1.2.5 are equivalent.

With $Y \in \{0,1\}$

$$\begin{aligned}
\mathbb{P}(Y=1|x,\vec{\beta}) = p \quad &= \frac{1}{1+e^{-x^T\vec{\beta}}} = s(x^T\vec{\beta}) \\
\mathbb{P}(Y=0|x,\vec{\beta}) = 1-p \quad &= \frac{1}{1+e^{x^T\vec{\beta}}} = s(-x^T\vec{\beta})
\end{aligned}$$

With $Y \in \{0,1\}$

$$\begin{aligned}
\mathbb{P}(Y=1|x,\vec{\beta}) = p \quad &= \frac{1}{1+e^{-x^T\vec{\beta}}} = s(x^T\vec{\beta}) \\
\mathbb{P}(Y=-1|x,\vec{\beta}) = 1-p \quad &= \frac{1}{1+e^{x^T\vec{\beta}}} = s(-x^T\vec{\beta})
\end{aligned}$$

or more compactly,

$$\mathbb{P}(Y=\pm 1|x,\vec{\beta}) = \frac{1}{1+e^{-yx^T\vec{\beta}}} = s(yx_i^T\vec{\beta})$$