1 Logistic regression

We can use the same type of machinery (as linear regression) to do classification. We have the same graphical model as in linear regressions, as below.

![Graphical model for logistic regression](image)

Figure 1: Graphical model for logistic regression (same as the graphical model for linear regression).

Problems of binary classification with linear regression (in which $y_n \sim N(\beta^T x, \sigma^2)$): (1) it will predict something other than 0 or 1, (2) a single outlier can affect greatly the model. (Note: In classification, $y_n$ is either zero or one; not drawn from Gaussian.)

Model $y$ as Bernoulli:

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

The parameters to the Bernoulli is a function for $x$. What $\mu$ should be used?

1. $\mu(x) = \beta^T x$: No, because $\mu(x)$ has to be within 0 and 1
2. $\mu(x) = \text{logistic}(\eta(x))$: maps $R \rightarrow (0, 1)$

**logistic function**: $\mu(x) = \frac{1}{1 - e^{-\eta(x)}}, \eta(x) = x^T \beta$

Note:

1. $\eta(x) \sim \infty, \mu(x) \sim 1$
2. \( \eta(x) \sim -\infty, \mu(x) \sim 0 \)

This specifies the model: \( y_n \sim Bernoulli(\mu(x)) \), where \( \mu(x) \) is defined above.

The logistic regression model implicitly places a "separating hyperplane" in the input space, and the conceptual line indicates where the probability to be 1/2 (for binary classification). (Only the closest data points matter, as in SVM)

The MLE of \( \beta \) focuses on the point near the boundary.

Finding the MLE of \( \beta \):

\[
\hat{\beta} = \arg \max_{\beta} \log p(y_{1..N}|x_{1..N}, \beta), \text{ where data are } \{(x_n, y_n)\}_{n=1}^N, y_n \in 0, 1
\]

\[
L = \log p(y_{1..N}|x_{1..N}, \beta)
\]

\[
= \sum_{n=1}^N \log p(y_n|x_n, \beta)
\]

\[
= \sum_{n=1}^N \log(\mu(x_n)^{y_n}(1 - \mu(x_n))^{1-y_n}) \quad \text{(We have suppressed the dependence on } \beta)\]

\[
= \sum_{n=1}^N y_n \log \mu(x_n) + (1 - y_n) \log(1 - \mu(x_n))
\]

First we calculate the derivative with respective to \( \beta_i \):

\[
\frac{dL_n}{d\beta_i} = \sum_{n=1}^N \frac{dL_n}{d\mu(x_n)} \frac{d\mu(x_n)}{d\beta_i}
\]

term#1: \( \frac{dL_n}{d\mu(x_n)} = \frac{y_n}{\mu(x_n)} - \frac{(1-y_n)}{1-\mu(x_n)} \)

term#2: \( \frac{d\mu(x_n)}{d\beta_i} = \frac{d\mu_n}{d\eta_n} \frac{d\eta_n}{d\beta_i} = \mu_n (1 - \mu_n) x_{ni} \)

Let \( \mu_n \) be \( \mu(x_n) = \frac{1}{1+e^{-\beta^T x_n}} \)

Let \( \eta_n \) be \( \log \frac{\mu_n}{1-\mu_n} \) (inverse of logistic function)

Then \( \frac{d\mu_n}{d\eta_n} = \mu_n (1 - \mu_n) \)

From the term#1 and term#2 above, we have:

\[
\frac{dL_n}{d\beta_i} = \sum_{n=1}^N \left( \frac{y_n}{\mu_n} - \frac{1-\mu_n}{1-\mu_n} \right) \mu_n (1 - \mu_n) x_{ni} = \sum_{n=1}^N (y_n - \mu_n) x_{ni}
\]

\[
E[y_n | x_n, \beta] = p(y_n = 1 | x_n, \beta) = \mu(x_n) = \mu_n, \text{ so } \frac{dL_n}{d\beta_i} = \sum_{n=1}^N (y_n - E[y_n | x_n, \beta]) x_{ni}
\]

Regression: \( L = \sum_{n=1}^N y_n \mu_n + (1 - y_n)(1 - \mu_n) + \|\beta\|_q \)

Connection to Naive Bayes:
Figure 2: Generative model.

Figure 3: Discriminative model.

Note: When you see more training data, you’ll see more outliers that might affect Naive Bayes, but not logistic regression or SVM.