Logistic regression

Daniel Hsu (COMS 4771)

The logistic regression model
Logistic regression is a model for binary classification data with feature vectors in \( \mathbb{R}^d \) and labels in \( \{-1, +1\} \). Data \((X_1, Y_1), \ldots, (X_n, Y_n)\) are treated as iid random variables taking values in \( \mathbb{R}^d \times \{-1, +1\} \), and for each \( x \in \mathbb{R}^d \),
\[
Y_i \mid X_i = x \sim \text{Bern}(\sigma(x^T w))
\]
where \( \sigma(t) = 1/(1 + \exp(-t)) \) is the sigmoid function. Here, \( w \in \mathbb{R}^d \) is the parameter of the model, and it is not involved in the marginal distribution of \( X_i \) (which we leave unspecified).

Maximum likelihood
The log-likelihood of \( w \) given \((x_1, y_1), \ldots, (x_n, y_n)\) is
\[
-\sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i^T w)) + \text{(terms that do not involve } w).\]
There is no closed-form expression for the maximizer of the log-likelihood. Nevertheless, we can approximately minimize the negative log-likelihood with gradient descent.

Empirical risk minimization
Maximum likelihood is very different from finding the linear classifier of smallest empirical zero-one loss risk. Finding the empirical zero-one loss risk minimizer is computationally intractable in general.

Finding a linear separator
There are special cases when finding the empirical zero-one loss risk minimizer is computationally tractable. One is when the training data is \textit{linearly separable}: i.e., when there exists \( w^* \in \mathbb{R}^d \) such that
\[
y_i x_i^T w^* > 0, \quad \text{for all } i = 1, \ldots, n.
\]

Claim. Define \( L(w) := \sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i^T w)) \). Suppose \((x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, +1\} \) is linearly separable. Then any \( \hat{w} \in \mathbb{R}^d \) with
\[
L(\hat{w}) < \inf_{w \in \mathbb{R}^d} L(w) + \ln(2)
\]
is a linear separator.

Proof. We first observe that the infimum\(^1\) (i.e., greatest lower bound) of \( L \) is zero. Let \( w^* \in \mathbb{R}^d \) be a linear separator, so \( s_i := y_i x_i^T w^* > 0 \) for all \( i = 1, \ldots, n \). For any \( r > 0 \),
\[
L(rw^*) = \sum_{i=1}^{n} \ln(1 + \exp(-rs_i)),
\]
\(^1\)https://en.wikipedia.org/wiki/Infimum_and_supremum
and therefore
\[ \lim_{r \to \infty} \sum_{i=1}^{n} \ln(1 + \exp(-rs_i)) = 0. \]

Every term \( \ln(1 + \exp(-y_i x_i^T w)) \) in \( L(w) \) is positive, so \( L(w) > 0 \). Therefore, we conclude that
\[ \inf_{w \in \mathbb{R}^d} L(w) = 0. \]

So now we just have to show that any \( \hat{w} \in \mathbb{R}^d \) with
\[ L(\hat{w}) < \ln(2) \]
is a linear separator. So let \( \hat{w} \) satisfy \( L(\hat{w}) < \ln(2) \), which implies
\[ \ln(1 + \exp(-y_i x_i^T \hat{w})) < \ln(2) \]
for every \( i = 1, \ldots, n \). Exponentiating both sides gives
\[ 1 + \exp(-y_i x_i^T \hat{w}) < 2. \]
Now subtracting 1 from both sides and taking logarithms gives
\[ -y_i x_i^T \hat{w} < 0. \]
This means that \( \hat{w} \) correctly classifies \( (x_i, y_i) \). Since this holds for all \( i = 1, \ldots, n \), it follows that \( \hat{w} \) is a linear separator.

\[ \square \]

**Surrogate loss**

Even if \( (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, +1\} \) is not linearly separable, approximately maximizing the log-likelihood can yield a good linear classifier. This is because maximizing \( L \) is the same as minimizing the **empirical logistic loss risk**:
\[ \hat{R}(w) := \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{log}}(y_i x_i^T w) \]
where
\[ \ell_{\text{log}}(z) := -\ln \sigma(z) \]
is the **logistic loss**. The logistic loss (up to scaling) turns out to be an upper-bound on the zero-one loss:
\[ \ell_{\text{zo}}(z) \leq \frac{1}{\ln 2} \ell_{\text{log}}(z), \]
where \( \ell_{\text{zo}}(z) = 1_{\{z \leq 0\}} \). If the empirical logistic loss risk is small, then the empirical zero-one loss is also small.

**Gradient descent for logistic regression**

The derivative of \( \ell_{\text{log}} \) is given by
\[
\frac{d\ell_{\text{log}}(z)}{dz} = -\frac{1}{\sigma(z)} \cdot \frac{d\sigma(z)}{dz} \\
= -\frac{1}{\sigma(z)} \cdot \sigma(z) \cdot \sigma(-z) \\
= -\sigma(-z).
\]
Therefore, by linearity and the chain rule, the negative gradient of $\hat{R}$ with respect to $w$ is

$$-\nabla \hat{R}(w) = -\frac{1}{n} \sum_{i=1}^{n} \nabla \ell_{\text{log}}(y_i x_i^T w)$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \left. \frac{d\ell_{\text{log}}(z)}{dz} \right|_{z=y_i x_i^T w} \cdot \nabla (y_i x_i^T w)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sigma(-y_i x_i^T w) \cdot y_i x_i.$$

Now suppose $A = [x_1 \cdots x_n]^T \in \mathbb{R}^{n \times d}$ and $b = [y_1 \cdots y_n]^T \in \mathbb{R}^n$. (Notice that we have omitted the $1/\sqrt{n}$ scaling that we had for least squares linear regression.) Then the negative gradient of $\hat{R}$ can be written as

$$-\nabla \hat{R}(w) = \frac{1}{n} A^T (b \circ \sigma(-b \circ (A w))),$$

where $u \circ v \in \mathbb{R}^n$ is the coordinate-wise product of vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$, and $\sigma(v) \in \mathbb{R}^n$ is the coordinate-wise application of the sigmoid function to $v \in \mathbb{R}^n$.

Gradient descent for logistic regression begins with an initial weight vector $w^{(0)} \in \mathbb{R}^d$, and then iteratively updates it by subtracting a positive multiple $\eta > 0$ of the gradient at the current iterate:

$$w^{(t)} := w^{(t-1)} - \eta \nabla \hat{R}(w^{(t-1)}).$$