Logistic regression

Sep 2022

Murphy chap8

Logistic regression

Recall that in QDA, decision boundaries between classes are generally parabolic.

When two classes have equal class conditional covariance, their decision boundary becomes linear.

LDA is a special case of QDA where we assume all classes share a common covariance, so all decision boundaries are linear.

The question is, what if we model the linear decision boundaries directly?

Assume that there are K classes, the goal is achieved if we define:

Logistic regression

$$\ln P(y=1|x) = \beta_1^T x - \ln Z,$$
$$\ln P(y=2|x) = \beta_2^T x - \ln Z,$$
$$\cdots\cdots$$
$$\ln P(y=K|x) = \beta_3^T x - \ln Z,$$

$\ln Z$ is added to ensure normalization:

$$\sum_{k=1}^{K} P(y=K|x) = 1.$$

Logistic regression

Exponentiating both sides

$$P(y = k|x) = \frac{1}{Z} e^{\beta_k^T x}, \ k = 1, ..., K$$

we see that

$$Z = \sum_{k=1}^{K} e^{\beta_k^T x}$$

$$P(y = k|x) = \frac{e^{\beta_k^T x}}{\sum_{k=1}^{K} e^{\beta_k^T x}}, \ k = 1, ..., K$$

Fit the model - cross entropy

$$D_{KL}(P||Q) = \int \frac{P(dx)}{Q(dx)} \log\left(\frac{P(dx)}{Q(dx)}\right) Q(dx),$$

$\frac{P(dx)}{Q(dx)}$ is the Radon-Nikodym derivative of P with respect to Q

If P and Q have densities, then

$$P(dx) = p(x)\mu(dx),\ Q(dx) = q(x)\mu(dx).$$

$$D_{KL}(P||Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) \mu(dx).$$

Fit the model - cross entropy

discrete form of the relative entropy of P from Q,

$$D_{KL}(P||Q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

Theorem 2. *Let P and Q be probabilities on a measurable space X. The following properties hold.*

(1) *$D_{KL}(P||Q) \geqslant 0$ with equality if and only if $P = Q$ as measures. If $\mu(dx)$ is any measure on X with respective to which P and Q have densities $p(x)$ and $q(x)$, equality holds if and only if $p(x) = q(x)$, $\mu(dx)$-a.e.*

(2) *$D_{KL}(P||Q)$ is jointly convex in (P, Q).*

(3) *In general $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.*

In practice, we write

$$\begin{aligned} D_{KL}(P||Q) &= -\sum_x p(x) \log\left(\frac{q(x)}{p(x)}\right) \\ &= -\sum_x p(x) \log q(x) + \sum_x p(x) \log p(x). \end{aligned}$$

Fit the model - cross entropy

discrete form of the relative entropy of P from Q,

$$D_{KL}(P||Q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

Theorem 2. *Let P and Q be probabilities on a measurable space X. The following properties hold.*

(1) *$D_{KL}(P||Q) \geqslant 0$ with equality if and only if $P = Q$ as measures. If $\mu(dx)$ is any measure on X with respective to which P and Q have densities $p(x)$ and $q(x)$, equality holds if and only if $p(x) = q(x)$, $\mu(dx)$-a.e.*

(2) *$D_{KL}(P||Q)$ is jointly convex in (P, Q).*

(3) *In general $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.*

Fit the model - MLE

Let $(x_i, y_i)_{i=1}^N$ be a set of observations.

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{N} P(y_i|x_i) \\
&= \prod_{i=1}^{N} \left(\prod_{k=1}^{K} P(y_i|x_i)^{\mu_k(y_i)} \right),
\end{aligned}
$$

where $\beta = (\beta_1, ..., \beta_K) \in \mathbb{R}^{d \times K}$, $y_i \in \{1, ..., K\}$ is the class label of the i-th obervation x_i and for each k, $c \mapsto \mu_k(c)$ if an indicator function over the classes, it equals 1 if $c = k$, equals 0 otherwise, i.e.

$$
\mu_k(c) = \begin{cases} 1, & c = k; \\ 0, & \text{otherwise.} \end{cases}
$$

The log-likelihood function is

$$
\log L(\beta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mu_k(y_i) \log P(y_i|x_i).
$$

Fit the model - MLE

$$\log L(\beta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mu_k(y_i) \log P(y_i|x_i)$$

The inner sum over classes is identified as the negative cross entropy of

$$\mu(y_i) \triangleq (\mu_1(y_i), ..., \mu_K(y_i)) \text{ from } p(y_i) \triangleq (P(y_i = 1|x_i), ..., P(y_i = K|x_i)),$$

i.e.

$$\log L(\beta) = -\sum_{i=1}^{N} H(\mu(y_i), p(y_i)).$$

Thus maximizing the likelihood function $\beta \mapsto \log L(\beta)$ is equivalent to minimizing the sum of cross entropies

$$\beta \mapsto -\sum_{i=1}^{N} H(\mu(y_i), p(y_i)).$$

Maximum entropy property

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} P(y_i = k|x_i)\log P(y_i = k|x_i).$$

The fixed-mean constraints,

$$\sum_{i=1}^{N} P(y_i = k|x_i)x_{ij} = \sum_{i=1}^{N} \mu_k(y_i)x_{ij},\ \forall k = 1, ..., K,\ j = 1, ..., p.$$

The normalization contraint,

$$\sum_{k=1}^{K} P(y_i = k|x_i) = 1,\ i = 1, ..., N.$$

Maximum entropy property

6.3. **Binary logistic regression.** For binary classification, we encode the class labels using $\{0, 1\}$. Note $P(y = 1|x) = 1 - P(y = 0|x)$. We may write

$$\ln P(y = 0|x) = \beta_0^T x - \ln Z,$$

$$\ln(1 - P(y = 0|x)) = \beta_1^T x - \ln Z.$$

Since we have incorporated the normalization condition, the coefficients β_0^T, β_1^T must be compatible in some sense. One way to include the compatibility is to write the model as (by substracting the second equation from the first),

$$\ln \frac{P(y = 0|x)}{1 - P(y = 0|x)} = \beta^T x, \ \ \beta \in \mathbb{R}^p.$$

The LHS as a function of $P(y = 0|x)$ is called the logit or log odds function

$$\operatorname{logit}(p) = \ln \frac{p}{1 - p},$$

which is the inverse of the logistic function

$$\sigma(t) = \frac{1}{1 + e^{-t}}.$$

Therefore

$$P(y = 0|x) = \sigma\left(\beta^T x\right) = \frac{1}{1 + e^{-\beta^T x}}.$$