Постановка задачи классификации

Задана $D_m = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ — выборка объёма m, где $y \sim \mathsf{P}(y|\boldsymbol{x})$. Найдя $\mathsf{P}(y|\boldsymbol{x})$, можно для каждого $\boldsymbol{x}$ давать предсказания значения y следующим способом

$$\hat{y} = \underset{y \in \{-1,1\}}{\operatorname{argmax}} \mathsf{P}(y|\boldsymbol{x})$$

Чтобы восстановить $\mathsf{P}(y|\boldsymbol{x})$ введём семейство функций $\mathcal{F} = \{f(\boldsymbol{x}, \boldsymbol{w}|\boldsymbol{w} \in \mathbb{R}^n, x \in \mathbb{R}^n,)\}$ таких, что $f(\boldsymbol{x}, \boldsymbol{w}) = p(y|\boldsymbol{x})$. Задана логистическая регрессия, поэтому

$$f(\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1 + \exp(-\boldsymbol{x}^T \boldsymbol{w})}$$

а вектор оптимальных параматеров $\boldsymbol{w}$ вычисляется как доставляющий максимум правдоподобию данных. Совместное правдоподобие:

$$P(D_m, \boldsymbol{w}) = p(D_m|\boldsymbol{w}) p(\boldsymbol{w})$$

где $p(\boldsymbol{w})$ — априорное распределние векторов параметров, которое является нормальным $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{A})$. Можно вычислить вектор оптимальных параметров $\hat{\boldsymbol{w}}$

$$\mathcal{L}(D_m, \boldsymbol{w}) = \sum_{i=1}^m y_i \ln P(y_i|\boldsymbol{x_i}, \boldsymbol{w}) + \ln p(\boldsymbol{w}) = -\sum_{i=1}^m y_i \ln(1 + \exp(-\boldsymbol{x}_i^T \boldsymbol{w})) + \frac{1}{2C}||\boldsymbol{w}||_2^2 \to \max_{\boldsymbol{w} \in \mathbb{R}^n}$$

Для неизвестного объекта $\boldsymbol{x}$ предсказываем целевую переменную следующим образом

$$y = \arg\max_{y \in \{0,1\}} \frac{1}{1 + \exp(-\boldsymbol{x}^T \hat{\boldsymbol{w}})}$$

Постановка задачи определения оптимального объёма выборки

Предположим задано m объектов выборки D, порождённой распределением $\mathsf{P}(\boldsymbol{x}, y)$. Объём m заранее недостаточен для обучения устойчивой логистической регресии. Необходимо, имея m объектов, спрогнозировать такое m^*, что каждый новый объект после m^* не будет давать новой информации о распределении параметров логистической регрессии, обученной на m^* объектах. Обозначим $\mathbb{P}(\boldsymbol{w}|m)$ распределение весов классификатора, обученного на выборке D_m порождённой $\mathbb{P}(\boldsymbol{x}, y)$. Задача прогнозирования оптимального объёма выборки имеет вид

$$m^* = \min m \in \mathbb{N} \; \forall k \in \mathbb{N} \to \rho(\mathsf{P}(\boldsymbol{w}|m), \mathsf{P}(\boldsymbol{w}|m+k)) < \varepsilon$$

где ρ — некоторая функция расстояния между распределениями, а ε — заданный порог.

Базовый метод

В качестве расстояния между распределениями мы используем расстояние Кульбака-Лейблера. Предполагая, что выборка простая и поскольку целевая переменная порождается распределением Бернулли, распределение векторов параметров является нормальным [ссылка Стрижов]. Удаляя из выборки по l элементов, получим набор подвыборок размера $m - l$. Обучив на этих подвыборках логистические регрессии, получим набор оптимальных векторов параметров. Этот набор порождён нормальным распределнием. Выбрав два

различных объёма m_1 и m_2, мы генерируем два нормальных распределния векторов парамтеров с матожиданиями $\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2$ и матрицами ковариации $\boldsymbol{A}_1, \boldsymbol{A}_2$. Таким образом

$$\rho(\mathsf{P}(\boldsymbol{w}|m), \mathsf{P}(\boldsymbol{w}|m+k)) = KL(\mathcal{N}(\hat{\boldsymbol{w}}_1, \boldsymbol{A}_1)||\mathcal{N}(\hat{\boldsymbol{w}}_2, \boldsymbol{A}_2)) =$$

$$\frac{1}{2}\left(tr(\boldsymbol{A}_2^{-1}\boldsymbol{A}_1) + (\hat{\boldsymbol{w}}_1 - \hat{\boldsymbol{w}}_2)^T\boldsymbol{A}_2^{-1}(\hat{\boldsymbol{w}}_1 - \hat{\boldsymbol{w}}_2) - n + \ln\left(\frac{\det(\boldsymbol{A}_2)}{\det(\boldsymbol{A}_1)}\right)\right)$$

Чтобы оценить оптимальный объём выборки поочерёдно добавляем к выборке по 1 элементу пока объём выборки не достигнет m. из выборки меняя l от $m-k$ до 1. Получим $m-k$ распределений $\mathsf{P}(\boldsymbol{w}|m-l)$. Чем больше $m-l$, тем точнее логистическая регрессия восстанавливает распределние $\mathsf{P}(y|\boldsymbol{x})$, поскольку обучена на большем количестве объектов. Обозначи m^* оптимальный объём выборки, т. е. такой, что по такой выборке распределение $\mathsf{P}(y|\boldsymbol{x})$ восстанавливается точно. Чем меньше расстояние $KL(\mathsf{P}(\boldsymbol{w}|m-l)||\mathsf{P}(\boldsymbol{w}|m^*)$, тем ближе $m-l$ к оптимальному размеру выборки. Поскольку m^* — оптимальный объём, $KL(\mathsf{P}(\boldsymbol{w}|m^*+p)||\mathsf{P}(\boldsymbol{w}|m^*)$ не меняется существенно при $p \to \infty$. Таким образом график зависимости $KL(\mathsf{P}(\boldsymbol{w}|m)||\mathsf{P}(\boldsymbol{w}|m^*)$ от m выходит на плато на m^*. По полученным $m-k$ распределниям аппроксимируем эту зависимость. Имея аппроксимирующую функцию, можно найти m^* как точку, на которой график функции выходит на плато.

*ЗАМЕТКА чтобы получить прогноз необходимого объёма, так или иначе, нужно разбираться с тем, как ведет себя матожидание и матрица ковариации как случайная величина, зависящая от размера выборки. Идея в том, что мы эмперически попробуем аппроксимировать функцию w от m, матрицу ковариации.