Машинное обучение, ФКН ВШЭ
Семинар №5

1 Предсказание вероятностей

Разберемся, каким требованиям должен удовлетворять классификатор, чтобы его выход можно было расценивать как оценку вероятности класса.

Пусть в каждой точке $x \in \mathbb{X}$ пространства объектов задана вероятность $p(y = +1 \,|\, x)$ того, что данный объект относится к классу $+1$, и пусть алгоритм $b(x)$ возвращает числа из отрезка $[0, 1]$. Потребуем, чтобы эти предсказания пытались в каждой точке x приблизить вероятность положительного класса $p(y = +1|x)$.

Разумеется, выполнение этого требования зависит от функции потерь — минимум ее матожидания в каждой точке x должен достигаться на данной вероятности:

$$\arg\min_{b \in \mathbb{R}} \mathbb{E}\left[L(y, b)|x\right] = p(y = +1|x).$$

Задача 1.1. *Покажите, что квадратичная функция потерь $L(y, b) = ([y = +1] - b)^2$ позволяет предсказывать корректные вероятности.*

Решение. Заметим, что поскольку алгоритм возвращает числа от 0 до 1, то его ответ должен быть близок к единице, если объект относится к положительному классу, и к нулю — если объект относится к отрицательному классу.

Запишем матожидание функции потерь в точке x:

$$\mathbb{E}\left[L(y, b)|x\right] = p(y = +1|x)(b - 1)^2 + (1 - p(y = +1|x))(b - 0)^2.$$

Продифференцируем по b:

$$\frac{\partial}{\partial b}\mathbb{E}\left[L(y, b)|x\right] = 2p(y = +1|x)(b - 1) + 2(1 - p(y = +1|x))b = 2b - 2p(y = +1|x) = 0.$$

Легко видеть, что оптимальный ответ алгоритма действительно равен вероятности:

$$b = p(y = +1|x).$$

$\blacksquare$

Задача 1.2. *Покажите, что абсолютная функция потерь $L(y, b) = |[y = +1] - b|$, $b \in [0; 1]$, не позволяет предсказывать корректные вероятности.*

Решение. Запишем матожидание функции потерь в точке x:

$$\begin{aligned}\mathbb{E}\left[L(y,b)|x\right] =& p(y=+1|x)|1-b| + (1-p(y=+1|x))|b| = \\ &= p(y=+1|x)(1-b) + (1-p(y=+1|x))b.\end{aligned}$$

Продифференцируем по b:

$$\frac{\partial}{\partial b}\mathbb{E}\left[L(y,b)|x\right] = 1 - 2p(y=+1|x) = 0.$$

Рассмотрим 2 случая:

1. $p(y=+1|x) = \frac{1}{2}$. Тогда $\mathbb{E}\left[L(y,b)|x\right] = \frac{1}{2} \quad \forall b \in [0;1]$, а потому классификатор не позволяет предсказывать корректную вероятность в точке x.

2. $p(y=+1|x) \neq \frac{1}{2}$. В этом случае интервал $(0;1)$ не содержит критических точек, а потому минимум матожидания достигается на одном из концов отрезка $[0;1]$:

$$\begin{aligned}\min_{b\in[0;1]} \mathbb{E}\left[L(y,b)|x\right] = \min\left(\mathbb{E}\left[L(y,0)|x\right], \mathbb{E}\left[L(y,1)|x\right]\right) = \\ \min\left(p(y=+1|x), 1-p(y=+1|x)\right).\end{aligned}$$

   Отсюда $\arg\min_{b\in[0;1]} \mathbb{E}\left[L(y,b)|x\right] \in \{0,1\}$, а потому классификатор также не позволяет предсказывать корректную вероятность в точке x.

∎

2 Задача на SVM

Задача 2.1. *Пусть $(w, b, \xi_1, \ldots, \xi_\ell)$ — оптимальное решение прямой задачи SVM. Предположим, что $\xi_3 > 0$. Выразите отступ объекта x_3 для обученного линейного классификатора через значения $(\xi_1, \ldots, \xi_\ell)$.*

Решение.

∎

3 Связь между l_2-регуляризацией и ранним остановом

Рассмотрим некоторую функцию потерь $L(w)$, как функцию от параметров линейной модели w. Разложим функцию $L(w)$ в окрестности локального минимума w^*:

$$\widehat{L}(w) = L(w^*) + \frac{1}{2}(w-w^*)^T H(w-w^*), \tag{3.1}$$

где H – Гессиан функции $L(w)$ относительно параметров w. В предположении, что w^* – локальный минимум, матрица H является положительно полуопределённой. Рассмотрим оптимизацию полученной функции $\widehat{L}(w)$ с помощью градиентного спуска. Градиент нашей аппроксимации:

$$\nabla\widehat{L}(w) = H(w - w^*). \tag{3.2}$$

Тогда формула пересчёта весов:

$$\begin{align}
w^{(\tau)} &= w^{(\tau-1)} - \varepsilon\nabla\widehat{L}(w^{(\tau-1)}) \tag{3.3}\\
&= w^{(\tau-1)} - \varepsilon H(w^{(\tau-1)} - w^*) \tag{3.4}\\
w^{(\tau)} - w^* &= (I - \varepsilon H)(w^{(\tau-1)} - w^*). \tag{3.5}
\end{align}$$

Так как матрица H положительно полуопределённая, мы можем переписать её в виде $H = Q\Lambda Q^T$, где Λ – диагональная матрица с собственными значениями на диагонали, а матрица Q составлена из ортонормированных собственных векторов.

$$\begin{align}
w^{(\tau)} - w^* &= (I - \varepsilon Q\Lambda Q^T)(w^{(\tau-1)} - w^*) \tag{3.6}\\
Q^T(w^{(\tau)} - w^*) &= (I - \varepsilon\Lambda)Q^T(w^{(\tau-1)} - w^*) \tag{3.7}
\end{align}$$

Рассмотрим ещё один шаг градиентного спуска:

$$\begin{align}
Q^T(w^{(\tau+1)} - w^*) &= (I - \varepsilon\Lambda)Q^T(w^{(\tau)} - w^*) \tag{3.8}\\
Q^T(w^{(\tau+1)} - w^*) &= (I - \varepsilon\Lambda)(I - \varepsilon\Lambda)Q^T(w^{(\tau-1)} - w^*) \tag{3.9}\\
Q^T(w^{(\tau+1)} - w^*) &= (I - \varepsilon\Lambda)^2 Q^T(w^{(\tau-1)} - w^*) \tag{3.10}\\
& \tag{3.11}
\end{align}$$

Если предположить, что мы стартовали из точки $w^{(0)} = 0$, то получаем формулу для значения весов после τ итераций:

$$Q^T w^{(\tau)} = (I - (I - \varepsilon\Lambda)^\tau)Q^T w^* \tag{3.12}$$

Теперь рассмотрим ту же самую функцию потерь $L(w)$ и ту же аппроксимацию, но добавим к аппроксимации l_2-регуляризацию:

$$\tilde{L}(w) = L(w^*) + \frac{1}{2}(w - w^*)^T H(w - w^*) + \frac{\alpha}{2}w^T w. \tag{3.13}$$

Запишем необходимое условие того, что точка $\tilde{w}$ является локальным минимумом:

$$\begin{align}
&\alpha\tilde{w} + H(\tilde{w} - w^*) = 0 \tag{3.14}\\
&(H + \alpha I)\tilde{w} = Hw^* \tag{3.15}\\
&\tilde{w} = (H + \alpha I)^{-1}Hw^* \tag{3.16}
\end{align}$$

Аналогично предыдущему случаю, запишем матрицу H через базис из ортонормированных собственных веторов $H = Q\Lambda Q^T$:

$$\tilde{w} = (Q\Lambda Q^T + \alpha I)^{-1} Q\Lambda Q^T w^* \tag{3.17}$$

$$= \left[Q(\Lambda + \alpha I)Q^T\right]^{-1} Q\Lambda Q^T w^* \tag{3.18}$$

$$= Q(\Lambda + \alpha I)^{-1} \Lambda Q^T w^* \tag{3.19}$$

$$Q^T \tilde{w} = (\Lambda + \alpha I)^{-1} \Lambda Q^T w^* \tag{3.20}$$

$$Q^T \tilde{w} = (I - (\Lambda + \alpha I)^{-1} \alpha) Q^T w^* \tag{3.21}$$

Выражение 3.12 можно сравнить с выражением 3.21. Видно, что $\tilde{w} = w^{(\tau)}$, если верно равенство

$$(I - \varepsilon\Lambda)^\tau = (\Lambda + \alpha I)^{-1}\alpha. \tag{3.22}$$

Таким образом, с точностью до выбора гиперпараметров, l_2-регуляризация эквивалентна раннему останову алгоритма оптимизации, в условия квадратичной аппроксимации. Более того, мы можем взять логарифм от одного из равенств из выражения 3.22 и получить следующее выражение

$$\tau \log(1 - \varepsilon\lambda_i) = -\log(1 + \lambda_i/\alpha). \tag{3.23}$$

Если у нас достаточно большой коэффициент регуляризации ($\lambda_i/\alpha \ll 1$) и не большой шаг градиентного спуска ($\varepsilon\lambda_i \ll 1$), то мы можем воспользоваться разложением $\log(1 + x)$ в окрестности нуля и записать

$$\tau \simeq \frac{1}{\varepsilon\alpha}. \tag{3.24}$$

То есть количество пройденных итераций обратно пропорционально коэффициенту регуляризации.