Машинное обучение, ФКН ВШЭ
Семинар №15

§1.1 Линейный Дискриминант Фишера

В этом семинаре будем рассматривать следующий подход к классификации объектов: выберем прямую в пространстве признаков с направляющим вектором w и спроецируем объект x на нее; если значение проекции окажется больше порога $-b$, то отнесем объект к классу $+1$, иначе к классу -1. Таким образом, классификатор будет иметь вид $a(x) = \operatorname{sign}(\langle w, x\rangle + b)$. Обучение классификатора сводится к поиску проекционной прямой. Будем выбирать ее так, чтобы после проецирования разброс точек из одного класса был как можно меньше, а расстояние между центрами классов было как можно больше. Формализуем эти требования. Обозначим через m_k центр k-го класса, $k \in Y$:

$$m_k = \frac{1}{N_k} \sum_{i:y_i=k} x_i.$$

Пусть s_k^2 — внутриклассовая дисперсия класса k:

$$s_k^2 = \sum_{i:y_i=k} (w^T x_i - w^T m_k)^2.$$

В качестве меры «сгруппированности» точек внутри своих классов возьмем сумму внутриклассовых дисперсий $s_{-1}^2 + s_{+1}^2$. В качестве меры расстояния между центрами проекций классов («межклассовой дисперсии») возьмем квадрат расстояния между этими центрами: $(w^T m_{-1} - w^T m_{+1})^2$. Чтобы совместить минимизацию первой величины и максимизацию второй, возьмем в качестве функционала их отношение. Получим следующую оптимизационную задачу:

$$J(w) = \frac{(w^T m_{-1} - w^T m_{+1})^2}{s_{-1}^2 + s_{+1}^2} \to \max_w.$$

Распишем данный функционал:

$$
\begin{aligned}
J(w) &= \frac{(w^T m_{-1} - w^T m_{+1})^2}{s_{-1}^2 + s_{+1}^2} = \\
&= \frac{\left(w^T\left(m_{-1} - m_{+1}\right)\right)^2}{\sum_{i:y_i=-1}\left(w^T\left(x_i - m_{-1}\right)\right)^2 + \sum_{i:y_i=+1}\left(w^T\left(x_i - m_{+1}\right)\right)^2} = \\
&= \frac{w^T(m_{-1} - m_{+1})(m_{-1} - m_{+1})^T w}{\sum_{i:y_i=-1} w^T(x_i - m_{-1})(x_i - m_{-1})^T w + \sum_{i:y_i=+1} w^T(x_i - m_{+1})(x_i - m_{+1})^T w} = \\
&= \frac{w^T(m_{-1} - m_{+1})(m_{-1} - m_{+1})^T w}{w^T\left(\sum_{i:y_i=-1}(x_i - m_{-1})(x_i - m_{-1})^T + \sum_{i:y_i=+1}(x_i - m_{+1})(x_i - m_{+1})^T\right) w}.
\end{aligned}
$$

Введем обозначения для ковариационных матриц:

$$
\begin{aligned}
&S_b = (m_{-1} - m_{+1})(m_{-1} - m_{+1})^T; \\
&S_w = \sum_{i:y_i=-1}(x_i - m_{-1})(x_i - m_{-1})^T + \sum_{i:y_i=+1}(x_i - m_{+1})(x_i - m_{+1})^T.
\end{aligned}
$$

Тогда функционал примет вид

$$
J(w) = \frac{w^T S_b w}{w^T S_w w} \to \max_w .
$$

Нам понадобится следующее правило векторного дифференцирования.

Задача 1.1. *Покажите, что если $f : \mathbb{R}^d \to \mathbb{R}$ и $g : \mathbb{R}^d \to \mathbb{R}$ — вещественные функции, то*

$$
\nabla_x \frac{f(x)}{g(x)} = \frac{g(x)\nabla_x f(x) - f(x)\nabla_x g(x)}{g^2(x)}.
$$

Воспользуемся полученным правилом, чтобы вычислить градиент функционала $J(w)$ и приравнять его нулю:

$$
\begin{aligned}
\nabla_w J(w) &= \frac{(S_b + S_b^T)w(w^T S_w w) - (S_w + S_w^T)w(w^T S_b w)}{(w^T S_w w)^2} = \\
&= 2\frac{S_b w(w^T S_w w) - S_w w(w^T S_b w)}{(w^T S_w w)^2} = \\
&= 0.
\end{aligned}
$$

Приходим к уравнению

$$
S_b w(w^T S_w w) = S_w w(w^T S_b w). \tag{1.1}
$$

Пусть минимум функционала $J(w)$ достигается на векторе w_*. Тогда этот вектор удовлетворяет уравнению (1.1). Поскольку классификатор (**??**) зависит только от направления вектора w и не зависит от его длины, мы можем проигнорировать скалярные множители. Получаем:

$$\begin{aligned} S_w w_* = & \\ & = \underbrace{\frac{w_*^T S_w w_*}{w_*^T S_b w_*}}_{\in\mathbb{R}} S_b w_* \propto \\ & \propto S_b w_* = \\ & = (m_{-1} - m_{+1}) \underbrace{(m_{-1} - m_{+1})^T w_*}_{\in\mathbb{R}} \propto \\ & \propto (m_{-1} - m_{+1}). \end{aligned}$$

Значит,

$$w_* = S_w^{-1}(m_{-1} - m_{+1}).$$

Далее в курсе будет рассмотрен нормальный дискриминантный анализ. Оказывается, что линейный дискриминант Фишера приводит к такому же вектору весов w, который может быть получен при нормальном дискриминантном анализе в предположении о равенстве ковариационных матриц классов.

§1.2 Ядровой Дискриминант Фишера

Рассмотрим ядровую версию линейного дискриминанта фишера. Обозначим отображение объекта x в спрямляющее пространство как $\varphi(x)$. Будем искать веса линейного классификатора в виде:

$$w = \sum_{i=1}^{\ell} \alpha_i \varphi(x_i).$$

Тогда, аналогично выражениям для ЛДФ, получаем:

$$m_k = \frac{1}{N_k} \sum_{i:y_i=k} \varphi(x_i),$$

$$w^T m_k = \frac{1}{N_k} \sum_{i:y_i=k} \sum_j \alpha_j \varphi(x_j)^T \varphi(x_i) = \frac{1}{N_k} \sum_{i:y_i=k} \sum_j \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle.$$

Введем обозначение матрицы Грама в спрямляющем пространстве $K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$, а также one-hot encode вектора для объектов k-го класса $\mathbf{1}_k$. Тогда выражение $w^T m_k$ примет вид:

$$w^T m_k = \frac{1}{N_k} \mathbf{1}_k^T K \alpha = \frac{1}{N_k} \alpha^T K \mathbf{1}_k.$$

Выражение для внутриклассовой дисперсии запишется следующим образом:

$$
\begin{aligned}
s_k^2 &= \sum_{i:y_i=k} (w^T\varphi(x_i) - w^T m_k)^2 = \sum_{i:y_i=k} w^T(\varphi(x_i) - m_k)(\varphi(x_i) - m_k)^T w \\
&= \sum_{i:y_i=k} \left[w^T\varphi(x_i)\varphi(x_i)^T w - 2w^T\varphi(x_i) m_k^T w + w^T m_k m_k^T w \right] \\
&= \sum_{i:y_i=k} \left[\left(\sum_j \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle \right)^2 - 2 \left(\sum_j \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle \right) m_k^T w + w^T m_k m_k^T w \right] \\
&= \sum_{i:y_i=k} \left(\sum_j \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle \right)^2 - N_k w^T m_k m_k^T w
\end{aligned}
$$

Обозначим $\mu_k = K\mathbf{1}_k$, тогда

$$
\begin{aligned}
\sum_k s_k^2 = \alpha^T K K^T \alpha - \sum_k N_k w^T m_k m_k^T w = \alpha^T K K^T \alpha - \sum_k \frac{1}{N_k} \alpha^T K \mathbf{1}_k \mathbf{1}_k^T K \alpha \\
= \alpha^T \left[KK - \sum_k \frac{1}{N_k} K \mathbf{1}_k \mathbf{1}_k^T K \right] \alpha = \alpha^T \left[KK - \sum_k \frac{1}{N_k} \mu_k \mu_k^T \right] \alpha
\end{aligned}
$$

Будем оптимизировать точно такой же функционал как и в случае с ЛДФ, только в спрямляющем пространстве

$$
J(w) = \frac{\alpha^T (\mu_{-1} - \mu_{+1})(\mu_{-1} - \mu_{+1})^T \alpha}{\alpha^T \left[KK - \frac{1}{N_{-1}} \mu_{-1} \mu_{-1}^T - \frac{1}{N_{+1}} \mu_{+1} \mu_{+1}^T \right] \alpha} \to \max_\alpha .
$$

Можно заметить, что функционал получился идентичный с точностью до обозначений, поэтому можно сразу выписать ответ

$$
\alpha_* = \left[KK - \frac{1}{N_{-1}} \mu_{-1} \mu_{-1}^T - \frac{1}{N_{+1}} \mu_{+1} \mu_{+1}^T \right]^{-1} (\mu_{-1} - \mu_{+1}).
$$