Третий семинар.
ММП, осень 2012–2013
2 октября

Темы семинара:

- Байесовская классификация;
- Нормальный дискриминантный анализ;
- Линейный дискриминант Фишера.

1 Первая половина

Начнем с короткого резюме вероятностной постановки, среднего риска классификатора и байесовского решающего правила

$$a(x) = \arg\max_{y \in \mathbb{Y}} \lambda_y P_y p(x|y).$$

Это случай, когда штраф зависит только от истинной классификации объекта. Коротко вспоминаем, что такое апостериорная вероятность и формулируем байесовский классификатор в его терминах.

Есть много способов оценивать плотность классов, сегодня будем говорить о нормальном дискриминантном анализе: мы будем полагать, что все классы имеют многомерное нормальное распределение. Это пример параметрического подхода к оценке плотности.

Задача 1. *Объекты двух классов $Y = \{-1, +1\}$ описываются точкой действительной оси $X = \mathbb{R}$. Априорные вероятности классов равны $P(Y = +1) = P(Y = -1) = 0.5$, величины потерь при ошибочной классификации одинаковы $\lambda_{+1,-1} = \lambda_{-1,+1}$, а при правильной классификации равны нулю $\lambda_{+1,+1} = \lambda_{-1,-1} = 0$. $p_1(x), p_2(x)$ — функции правдоподобия классов $+1$ и -1 соответственно. Выписать формулу Байесовского классификатора в явном виде, если:*

1. *$p_1(x)$ и $p_2(x)$ — гладкие функции, не касающиеся друг друга ни в одной точке, имеющие ровно n точек пересечения: $x_1 < x_2 < \cdots < x_n$.*

2. *$p_1(x) \equiv \mathcal{N}(0, 1)$, а $p_2(x) \equiv \mathcal{N}(0, 2z)$, где z — квантиль стандартного нормального распределения уровня 95%. Вычислите также байесовский уровень ошибки.*

Дальше вкратце вспоминаем, что такое декореллирующее преобразование координат, и как выглядят линии уровня многомерного нормального распределения.

Задача 2. *Каждый из двух классов имеет 2-мерное нормальное распределение со средними* $\boldsymbol{\mu_1} = \left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)$, $\boldsymbol{\mu_2} = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$ *и ковариационными матрицами* $\Sigma_1 = \Sigma_2 = \left(\begin{smallmatrix}2&1\\1&2\end{smallmatrix}\right)$. *Априорные вероятности классов* $P_1 = 1/3$, $P_2 = 2/3$. *Штрафы за ошибочную классификацию равны.*
а) Найдите оси симметрии линий уровня любой из функции правдоподобия класов.
б) Запишите функцию правдоподобия в новых координатах. Что это за преобразование? Какие полезные свойства вы в нем видите?
в) Выпишите аналитическую форму разделяющей поверхности байесовского классификатора в новых декореллированных координатах. Запишите поверхность в исходных координатах.

Решение. *а)* $\left(\begin{smallmatrix}2&1\\1&2\end{smallmatrix}\right) = \begin{pmatrix}1/\sqrt{2}&1/\sqrt{2}\\1/\sqrt{2}&-1/\sqrt{2}\end{pmatrix}\left(\begin{smallmatrix}3&0\\0&1\end{smallmatrix}\right)\begin{pmatrix}1/\sqrt{2}&1/\sqrt{2}\\1/\sqrt{2}&-1/\sqrt{2}\end{pmatrix}$
б) Запишем ф.п. для второго класса:

$$p_2(x) = \frac{1}{2\pi\sqrt{3}}\exp\left\{-\frac{1}{2}(t-\mu')^\top \left(\begin{smallmatrix}1/3&0\\0&1\end{smallmatrix}\right)(t-\mu')\right\};$$
$$t = \begin{pmatrix}1/\sqrt{2}&1/\sqrt{2}\\1/\sqrt{2}&-1/\sqrt{2}\end{pmatrix}\left(\begin{smallmatrix}x_1\\x_2\end{smallmatrix}\right); \quad \mu' = \begin{pmatrix}1/\sqrt{2}&1/\sqrt{2}\\1/\sqrt{2}&-1/\sqrt{2}\end{pmatrix}\left(\begin{smallmatrix}1\\1\end{smallmatrix}\right) = \left(\begin{smallmatrix}\sqrt{2}\\0\end{smallmatrix}\right).$$

в)

$$\ln(1/3) + C - \frac{1}{2}\left(\frac{t_1^2}{3} + t_2^2\right) = \ln(2/3) + C - \frac{1}{2}\left(\frac{(t_1-\sqrt{2})^2}{3} + t_2^2\right),$$

или

$$t_1 = \frac{1 - 3\ln 2}{\sqrt{2}}.$$

2 Вторая половина

2.1 Другой взгляд на ЛДФ. (Bishop, p. 186)

В лекции К. В. Воронцова ЛДФ вводится с положения ковариационных матриц классов равными и последовательным ее оцениванием с помощью обучающей выборки $\{(x_i, y_i)\}_{i=1}^{\ell}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{Y}$:

$$\hat{\Sigma} = \frac{1}{\ell}\sum_{i=1}^{\ell}(x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^\top,$$

где $\hat{\mu}_y$ — среднее выборочное значение объектов из класса $y \in \mathbb{Y}$. Эту эвристику, оказывается, можно обосновать несколькими способами. Приведем одну из постановок задач, которая ведет к получению тех же результатов.

Рассмотрим задачу классификации с двумя классами $\mathbb{Y} = \{1, 2\}$ и попробуем решать ее с помощью линейной проекции объектов обучающей выборки на прямую: $t_i = w^\top x_i$, $w \in \mathbb{R}^d$. Выбрав порог w_0, мы можем принимать решение о классификации, сравнивая с ним t_i.

Очевидно, выбор w, основанный на максимальном удалении проекций средних выборочных значений классов, имеет недостатки. Предложим другой подход. Определим s_1^2 и s_2^2 — *внутриклассные дисперсии* проекций:

$$s_1^2 = \sum_{y_i=1}(t_i - m_1)^2; \quad s_2^2 = \sum_{y_i=2}(t_i - m_2)^2$$
$$m_1 = \frac{1}{\ell_1}\sum_{y_i=1} t_i \equiv w^\top \hat{\mu}_1; \quad m_2 = \frac{1}{\ell_1}\sum_{y_i=2} t_i \equiv w^\top \hat{\mu}_2.$$

Полной внутриклассной дисперсией назовем сумму $s_1^2 + s_2^2$. Нашей целью будет максимизировать отношение *межклассовой дисперсии* $(m_1 - m_2)^2$ и полной внутриклассовой:

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \to \max_{w \in \mathbb{R}^d}.$$

Распишем все величины с помощью w:

$$J(w) = \frac{\left(w^\top(\hat{\mu}_2 - \hat{\mu}_1)\right)^2}{\sum_{y_i=1}(w^\top x_i - w^\top \hat{\mu}_1)^2 + \sum_{y_i=2}(w^\top x_i - w^\top \hat{\mu}_2)^2} =$$
$$= \frac{w^\top(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^\top w}{\sum_{y_i=1} w^\top(x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^\top w + \sum_{y_i=2} w^\top(x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)^\top w} =$$
$$= \frac{w^\top(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^\top w}{w^\top\left(\sum_{y_i=1}(x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^\top + \sum_{y_i=2}(x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)^\top\right) w}.$$

Обозначим

$$S_b = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^\top;$$
$$S_w = \left(\sum_{y_i=1}(x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^\top + \sum_{y_i=2}(x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)^\top\right),$$

тогда

$$J(w) = \frac{w^\top S_b w}{w^\top S_w w}.$$

Нам остается приравнять производную того выражения по w нулю:

$$\frac{w^\top S_w w(S_b^\top w + \ S_b w) - w^\top S_b w(S_w^\top w + \ S_w w)}{(w^\top S_w w w)^2} = 0$$

Поскольку обе матрицы S_b и S_w симметричны, то приходим к уравнению

$$(w^\top S_w w) S_b w = (w^\top S_b w) S_w w.$$

Заметим, что $S_b w \propto \hat{\mu}_2 - \hat{\mu}_1$, поэтому приходим к

$$w \propto S_w^{-1}(\hat{\mu}_2 - \hat{\mu}_1).$$

Таким образом, максимизируя функционал $J(w)$, мы получили ту же самую нормаль разделяющей поверхности, как в предположении совпадающих ковариационных матриц в рамках нормального дискриминантного анализа.

2.2 ЛДФ vs МНК (Bishop, p. 189), (Tibshirani ex. 4.2)

Мы снова рассматриваем задачу классификации с $\mathbb{X} = \mathbb{R}^d$. Покажем, что в случае двух классов $\mathbb{Y} = \{1, 2\}$ между ЛДФ и МНК при особом выборе ответов на объектах обучающей выборки есть прямая связь.

МНК, как мы знаем, используется, как правило, в задачах восстановления регрессии и стремится уменьшить квадратичную ошибку на обучающей выборке. Рассмотрим следующую задачу регрессии, полученную из рассматриваемой задачи классификации: положим ответы t_i на объектах обучающей выборки класса 1 и 2 равными ℓ/ℓ_1 и $-\ell/\ell_2$ соответственно. Нашей целью будет поиск вектора весов $w \in \mathbb{R}^d$ и константы w_0, минимизирующих среднеквадратичный риск на обучающей выборке:

$$\frac{1}{2}\sum_{i=1}^{\ell}(w^\top x_i + w_0 - t_i)^2 \to \min_{w,w_0}.$$

Приравняв производные по w и w_0 нулю, получаем:

$$\sum_{i=1}^{\ell}(w^\top x_i + w_0 - t_i) = 0; \tag{1}$$

$$\sum_{i=1}^{\ell}(w^\top x_i + w_0 - t_i)x_i = 0. \tag{2}$$

Из первого равенства с учетом $\sum_i t_i = 0$ получаем $w_0 = -w^\top\mu$, где μ — среднее выборочное объектов и

$$\mu \equiv \frac{\ell_1}{\ell}\mu_1 + \frac{\ell_2}{\ell}\mu_2.$$

Поскольку

$$\sum_{i=1}^{\ell} t_i x_i = \frac{\ell}{\ell_1}\sum_{y_i=1} x_i - \frac{\ell}{\ell_2}\sum_{y_i=2} x_i = \ell(\mu_1 - \mu_2),$$

то второе равенство преобразуется к виду

$$\sum_{i=1}^{\ell}(w^\top x_i + w_0)x_i = \ell(\mu_1 - \mu_2).$$

Продолжим работу с левой частью последнего равенства:

$$\sum_{i=1}^{\ell}(w^\top x_i + w_0)x_i = \sum_{i=1}^{\ell}(w^\top x_i - w^\top\mu)x_i = \sum_{i=1}^{\ell} x_i(x_i - \mu)^\top w = Mw,$$

где $M \in \mathbb{R}^{d\times d}$ — некоторая матрица. Докажем, что $M = S_w + \frac{\ell_1\ell_2}{\ell}S_b$. Будем вести

цепочку неравенств в обратную сторону:

$$S_w + \frac{\ell_1\ell_2}{\ell} S_b = \sum_{y_i=1}(x_i - \mu_1)(x_i - \mu_1)^\top + \sum_{y_2=1}(x_i - \mu_2)(x_i - \mu_2)^\top + \frac{\ell_1\ell_2}{\ell} S_b =$$

$$= \sum_{i=1}^{\ell} x_i x_i^\top + \ell_1\mu_1\mu_1^\top + \ell_2\mu_2\mu_2^\top - \sum_{y_i=1} x_i\mu_1^\top - \sum_{y_i=1}\mu_1 x_i^\top - \sum_{y_i=2} x_i\mu_2^\top - \sum_{y_i=2}\mu_2 x_i^\top + \frac{\ell_1\ell_2}{\ell} S_b =$$

$$= \sum_{i=1}^{\ell} x_i x_i^\top + \ell_1\mu_1\mu_1^\top + \ell_2\mu_2\mu_2^\top - 2\ell_1\mu_1\mu_1^\top - 2\ell_2\mu_2\mu_2^\top + \frac{\ell_1\ell_2}{\ell} S_b =$$

$$= \sum_{i=1}^{\ell} x_i x_i^\top - \ell_1\mu_1\mu_1^\top - \ell_2\mu_2\mu_2^\top + \frac{\ell_1\ell_2}{\ell}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^\top =$$

$$= \sum_{i=1}^{\ell} x_i x_i^\top - \frac{\ell_1\ell_2}{\ell}(\mu_2\mu_1^\top + \mu_1\mu_2^\top) + \frac{\ell_1\ell_2 - \ell_2\ell}{\ell}\mu_2\mu_2^\top + \frac{\ell_1\ell_2 - \ell_1\ell}{\ell}\mu_1\mu_1^\top =$$

$$= \sum_{i=1}^{\ell} x_i x_i^\top - \frac{\ell_1\ell_2}{\ell}(\mu_2\mu_1^\top + \mu_1\mu_2^\top) - \frac{\ell_2^2}{\ell}\mu_2\mu_2^\top - \frac{\ell_1^2}{\ell}\mu_1\mu_1^\top =$$

$$= \sum_{i=1}^{\ell} x_i x_i^\top - \frac{1}{\ell}(\ell_1\mu_1 + \ell_2\mu_2)(\ell_1\mu_1^\top + \ell_2\mu_2^\top) = \sum_{i=1}^{\ell} x_i x_i^\top - (\ell_1\mu_1 + \ell_2\mu_2)\mu^\top =$$

$$= \sum_{i=1}^{\ell} x_i x_i^\top - \sum_{i=1}^{\ell} x_i\mu^\top = \sum_{i=1}^{\ell} x_i(x_i - \mu)^\top = M.$$

Таким образом, уравнение (2) переписывается в виде

$$(S_w + \frac{\ell_1\ell_2}{\ell} S_b)w = \ell(\mu_1 - \mu_2).$$

Снова, поскольку $S_b w \propto (\mu_2 - \mu_1)$, то получаем

$$w \propto S_w^{-1}(\mu_2 - \mu_1).$$

Таким образом, нормали к гиперплоскостям для ЛДФ и МНК для описанной задачи регрессии параллельны. Однако отметим, что свободные члены w_0 в общем случае не совпадают. Как правило, свободный член можно выбрать исходя из соображений хорошего риска на обучающей выборке. МНК не использует предположения о нормальности распределений классов, таким образом описанный подход можно применять и в более общих случаях, когда никакой нормальности нет.

3 Проверочная

1.а. Выведите оценку максимального правдоподобия для среднего одномерного нормального распределения.
1.б. Выведите оценку максимального правдоподобия для параметра p распределения Бернулли.
2.а. Вычислите дисперсию биномиального распределения.
2.б. Вычислите матожидание биномиального распределения.
3. Напишите формулу плотности n-мерного нормального распределения.

4 Домашнее задание

1. Пусть два класса имеют n-мерные нормальные распределения со средними $\boldsymbol{\mu_1}$ и $\boldsymbol{\mu_2}$ и одинаковыми ковариационными матрицами Σ. Докажите, что разделяющая поверхность байесовского классификатора проходит через середину отрезка, соединяющего средние значения двух классов, параллельно линиям уровня функций правдоподобия классов.

2. Два класса имеют двухмерные нормальные плотности с параметрами (μ_1, Σ_1) и (μ_2, Σ_2). Априорные вероятности классов — P_1 и P_2. Величины потерь при неправильной классификации — λ_1 и λ_2.

а) Найдите вид байесовской разделяющей поверхности, если $\Sigma_1 = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, $\Sigma_2 = \left(\begin{smallmatrix} 3 & -2 \\ -2 & 3 \end{smallmatrix}\right)$, $\mu_1 = \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$, $\mu_2 = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$, $P_1 = P_2 = \frac{1}{2}$, $\lambda_1 = \lambda_2$;

б) При каких параметрах задачи разделяющей поверхностью байесовского классификатора будет пара параллельных прямых?