Обучение с учителем. Дискриминантный анализ. Логистическая регрессия. Feature selection & extraction.

Владимир Агеев, 622 гр.

11 января 2018 г.

Содержание

1 Постановка задачи

Обозначим $\boldsymbol{\xi} \in \mathbb{R}^p$ – случайный вектор, вектор признаков, $\eta \in \mathcal{G} = \{G_k\}_{k=1}^K$ – дискретная случайная величина, отметка класса, $P(\boldsymbol{\xi}, \eta)$ – их совместное распределение.

Пусть нам дана выборка $(\mathbf{x}_i, y_i)_{i=1}^N$ – N реализаций случайных векторов $(\boldsymbol{\xi}, \eta)$. По выборке необходимо построить функцию $a : \mathbb{R}^p \to \mathcal{G}$, которая по реализации случайной величины $\boldsymbol{\xi}$ возвращает соответствующую метку класса $G \in \mathcal{G}$. В некотором смысле этот классификатор должен оказаться самым лучшим, минимально ошибаться.

Постановка задачи, в которой предполагается строить классификатор (или другую зависимость) исходя из данных ответов в выборке, называется обучением с учителем.

2 Байесовский классификатор

В качестве меры ошибки предсказания введем функцию потерь. Рассмотрим матрицу $\mathbf{L}$ размера $K \times K$, где $K = card(\mathcal{G})$. На диагонали $\mathbf{L}$ стоят нули, а $\mathbf{L}(i,j) = \lambda_{ij}$ – цена ошибки отнесения элемента класса G_i к классу G_j. Часто используется 0-1 функция потерь, где каждая ошибка оценивается единичкой.

Математическое ожидание функции потерь (средний риск):

$$R(a) = \mathbb{E}(\mathbf{L}(\eta, a(\boldsymbol{\xi}))) = \mathbb{E}_{\boldsymbol{\xi}} \sum_{k=1}^{K} L(G_k, a(\boldsymbol{\xi})) P(G_k \mid \boldsymbol{\xi}).$$

Отсюда строим функцию классификации

$$a(\mathbf{x}) = \arg\min_{G \in \mathcal{G}} \sum_{k=1}^{K} L(G_k, G) P(G_k \mid \boldsymbol{\xi} = \mathbf{x}).$$

Если подставим сюда 0-1 функцию потерь, получим

$$a(\mathbf{x}) = \arg\min_{G \in \mathcal{G}} 1 - P(G \mid \boldsymbol{\xi} = \mathbf{x}).$$

Или, что то же самое

$$a(\mathbf{x}) = \arg\max_{G \in \mathcal{G}} P(G \mid \boldsymbol{\xi} = \mathbf{x}) = \arg\max_{G \in \mathcal{G}} P(G) P(\boldsymbol{\xi} \mid \eta = G).$$

Это решение называется байесовским классификатором, а такой подход – принципом максимума апостериорной вероятности.

3 Задача классификации на языке ML

На языке ML рассмотрим задачу классификации. Пусть нам дана простая выборка $X^n = (x_i, y_i)_{i=1}^n$, в которой наблюдения лежат в одном из двух классов $Y = \{-1, +1\}$.

Введем некоторые понятия:

- $a(x, \beta) = \text{sign } f(x, \beta)$ – семейство классификаторов;
- $M_i(\beta) = y_i f(x_i, \beta)$ – отступ объекта x_i, мера принадлежности объекта x_i классу y_i;
- $\mathcal{L}(M_i(\beta))$ – монотонно невозрастающая функция потерь, мажорирующая 0-1 функцию потерь $[M < 0]$.

Задача поиска классификатора $a(x, \beta)$ сводится к задаче минимизации эмпирического риска

$$Q(\beta, \mathbf{X}) = \sum_{i=1}^{N} [M_i(\beta) < 0] \leq \sum_{i=1}^{N} \mathcal{L}(M_i(\beta)) \to \min_{\beta}.$$

Положив $\mathcal{L}(M_i(\theta)) = -\log P(x_i, y_i; \theta)$ получаем эквивалентность с задачей максимизации правдоподобия

$$\sum_{i=1}^{N} \log P(x_i, y_i; \theta) \to \max_{\theta}.$$

4 Дискриминантный анализ

Для построения байесовского классификатора, нам необходимо знать апостериорные вероятности $P(G \mid \boldsymbol{\xi} = \mathbf{x})$. Обозначим $p_k(\mathbf{x}) = P(\boldsymbol{\xi} = \mathbf{x} \mid \eta = G_k)$ условные плотности классов, $\pi_k = P(\eta = G_k)$ – априорные вероятности, $\sum_{k=1}^{K} \pi_k = 1$. По теореме Байеса получим

$$P(G = k \mid X = x) = \frac{p_k(x)\pi_k}{\sum_{i=1}^{K} p_i(x)\pi_i}.$$

Возникает вопрос: откуда брать априорные вероятности?

- Брать равновероятные: $\pi_i = \frac{1}{K}$;
- Брать пропорционально объемам классов $\pi_i = \frac{n_i}{N}$;
- Соответственно имеющейся информации. Например, исходя из цены ошибки классификации.

4.1 Квадратичный дискриминантный анализ

Предположим, что каждый класс имеет многомерное нормальное распределение $P(\boldsymbol{\xi} \mid \eta = G_k) = \mathcal{N}_p(\mu_k, \boldsymbol{\Sigma}_k)$, его плотность

$$p_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)}.$$

Подставим плотности в байесовский классификатор, который мы получили в предыдущем пункте и получим

$$a(\mathbf{x}) = \arg\max_{i\in 1\dots K} \pi_i p_i(\mathbf{x}) = \arg\max_{i\in 1\dots K} \log(\pi_i p_i(\mathbf{x})) = \arg\max_{i\in 1\dots K} \log(\pi_i) + \log(p_i(\mathbf{x})) =$$
$$= \arg\max_{i\in 1\dots K}(-\frac{1}{2}(\mathbf{x}-\mu_i)^{\mathrm{T}}\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mu_i) - \frac{1}{2}\log(|\mathbf{\Sigma}_i|) + \log(\pi_i)) = \arg\max_{i\in 1\dots K} g_i(\mathbf{x}).$$

Заметим, что получившийся классификатор квадратично зависит от $\mathbf{x}$, отсюда название quadratic discriminant analysis.

4.2 Линейный дискриминантный анализ

Предположим теперь, что классы имеют нормальное распределение с одинаковой ковариационной матрицей, то есть $P(\boldsymbol{\xi} \mid \eta = G_k) = \mathcal{N}_p(\mu_k, \Sigma)$. Отсюда следует, что классификатор, полученный в предыдущем пункте, можно упростить следующим образом:

$$a(\mathbf{x}) = \arg\max_{i\in 1\dots K}(-\frac{1}{2}(\mathbf{x}-\mu_i)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i) - \frac{1}{2}\log(|\mathbf{\Sigma}|) + \log(\pi_i)) =$$
$$= \arg\max_{i\in 1\dots K}(-\frac{1}{2}\mu_i^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mu_i + \mu_i^{\mathrm{T}}\mathbf{\Sigma}^{-1}x + \log(\pi_i)) = \arg\max_{i\in 1\dots K}\delta_i(x).$$

Такой классификатор зависит от $\mathbf{x}$ линейно.

Разделяющая два класса гиперплоскость определяется так

$$\{\mathbf{x} : \delta_i(\mathbf{x}) = \delta_j(\mathbf{x})\} =$$
$$= \{x : -\frac{1}{2}(\mu_i-\mu_j)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mu_i+\mu_j) + (\mu_i-\mu_j)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x} + \log(\pi_i/\pi_j) = 0\}.$$

От соотношения между априорными вероятностями зависит положение границы относительно классов (к какому она ближе).

4.3 Оценка параметров

На практике параметры распределений классов нам не известны, поэтому предлагается использовать следующие оценки максимального правдоподобия параметров нормальных плотностей классов.

- Среднее $\widehat{\mu}_i = \frac{1}{n_i}\sum\limits_{j:y_j=G_i}\mathbf{x}_j$,
- Ковариационная матрица класса $\widehat{\mathbf{\Sigma}}_i = \frac{1}{n_i-1}\sum\limits_{j:y_j=G_i}(\mathbf{x}_j-\widehat{\mu}_i)^{\mathrm{T}}(\mathbf{x}_j-\widehat{\mu}_i)$,
- Pooled ковариационная матрица $\widehat{\mathbf{\Sigma}} = \sum\limits_{j=1}^{K}\frac{n_i-1}{n-K}\widehat{\mathbf{\Sigma}}_i$.

4.4 Возвращение к вероятностям

Если нам необходимо получить вероятности отношения $\mathbf{x}$ к классу i, то, вычислив $\widehat{\delta}_i(\mathbf{x})$, можно вычислить

$$\widehat{P}(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x}) = \frac{e^{\widehat{\delta}_i(\mathbf{x})}}{\sum\limits_{j=1}^{K} e^{\widehat{\delta}_j(\mathbf{x})}}.$$

4.5 Regularized Discriminant Analysis

Оценка ковариационной матрицы $\widehat{\mathbf{\Sigma}}_i$ может оказаться выражденной или плохо обусловленной. Опишем компромис между LDA и QDA, а так же борьбу с мультиколлинеарностью.

- Regularized Discriminant Analysis. Рассматривается матрица $\widehat{\mathbf{\Sigma}}_i(\alpha) = \alpha\widehat{\mathbf{\Sigma}}_i + (1-\alpha)\widehat{\mathbf{\Sigma}}$, где $\widehat{\mathbf{\Sigma}}$ – pooled ковариационая матрица. Здесь $\alpha \in [0,1]$ порождает континуум моделей между LDA и QDA, выбирается скользящим контролем.
- Дополнительно к предыдущему методу можно похожим образом модифицировать pooled ковариационную матрицу и рассматривать $\widehat{\Sigma}(\gamma) = \gamma\widehat{\mathbf{\Sigma}} + (1-\gamma)\sigma^2\mathbf{I}_p$, где γ определяет вид ковариационной матрицы и выбирается скользящим контролем.

4.6 Наивный байесовский классификатор

Предположим, что признаки независимы внутри групп и имеют нормальное распределение

$$p_i(x) = \prod_{j=1}^{p} p_{ij}(x_j), \quad p_{ij}(x_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}^2}}.$$

Отсюда классифицирующую функцию можно представить в виде

$$\delta_i(x) = -\frac{1}{2}\sum_{j=1}^{p} \frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}^2} + \log(\pi_i).$$

Аналогично подходам выше, можно подбирать ковариационную матрицу скользящим контролем в виде

$$\widehat{\mathbf{\Sigma}}_i(\alpha) = \alpha\widehat{\mathbf{\Sigma}}_i + (1-\alpha)\ \mathrm{diag}(\sigma_{i1}^2, \ldots, \sigma_{ip}^2).$$

Такой подход может быть полезен, когда признаков очень много и оценивать плотности классов оказывается сложно. Плотности p_{ki} можно оценивать по отдельности, а если признак дискретный, для этого можно использовать гистограмму.

Не смотря на такое оптимистичное предположение, наивный Байес часто превосходит более сложные методы.

4.7 Уменьшение размерности

К feature extraction в традиционном дискриминантном анализе подходят следующим образом.

Задача: найти линейное преобразование $\mathbf{Z} = A^{\mathrm{T}}\mathbf{X}$, в результате которого получаются признаки наилучшим образом разделяющие группы. Хотелось бы, чтобы эти признаки оказались ортогональны. Далее опишем эту задачу более формально.

- $\widehat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, – pooled ковариационная матрица
- $\mathbf{X}^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^{\mathrm{T}}\mathbf{X}$ – нормировали матрицу данных относительно pooled ковариационной матрицы
- Перейдем от матрицы данных $\mathbf{X}^*$ к матрице центров групп $\mathbf{M}_{K,p}$
- Вычислим внутриклассовую ковариационную матрицу

$$\mathbf{W} = \frac{1}{n-K}\sum_{i=1}^{K}\sum_{j:y_j=G_i}(\mathbf{x}_j - \widehat{\mu}_i)^{\mathrm{T}}(\mathbf{x}_j - \widehat{\mu}_i)$$

- Вычисляем межклассовую ковариационную матрицу (с точностью до коэффициента)

$$\mathbf{B} = \sum_{i=1}^{K} n_i(\widehat{\mu}_i - \widehat{\mu})^{\mathrm{T}}(\widehat{\mu}_i - \widehat{\mu}).$$

Пусть $\zeta = A\boldsymbol{\xi}$ – новый признак, тогда распределение $P(\zeta \mid \eta = G_k) = \mathcal{N}_p(A^{\mathrm{T}}\mu_k, A^{\mathrm{T}}\mathbf{\Sigma}_k A)$.

На выборочном языке новые признаки $Z = A^{\mathrm{T}}\mathbf{X}$. Выборочная ковариационная матрица (с точностью до коэффициента) новых признаков имеет вид

$$A^{\mathrm{T}}\mathbf{T}A = A^{\mathrm{T}}(\mathbf{W} + \mathbf{B})A = A^{\mathrm{T}}\mathbf{W}A + A^{\mathrm{T}}\mathbf{B}A,$$

где $\mathbf{T}$ – total covariance matrix, первое слагаемое – оценка внутригрупповых отклонений, а второе – оценка межгрупповых отклонений. Воспользовавшись критерием Фишера перейдем к обобщенной задаче на собственные числа и собственные вектора:

$$\frac{A^{\mathrm{T}}\mathbf{B}A}{A^{\mathrm{T}}\mathbf{W}A} \to \max_A.$$

Путь $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ – собственные числа матрицы $\mathbf{W}^{-1}\mathbf{B}$, а $A_1, \ldots, A_d$ – соответствующие им собственные вектора. Тогда максимум выше равен λ_1 и достигается на A_1. При этом $A_i^{\mathrm{T}}\mathbf{W}A_j = 0$. Далее

$$\max_{A, A\perp A_1} \frac{A^{\mathrm{T}}\mathbf{B}A}{A^{\mathrm{T}}\mathbf{W}A} = \lambda_2,$$

достигается на A_2 и так далее.

Вектора A_i называют каноническими коэффициентами, а новые признаки Z_i – каноническими переменными, Z_i ортогональны в обычном смысле.

4.8 Значимость канонических переменных

Возникает вопрос: сколько канонических переменных нам окажется достаточно взять? Другими словами, нужно проверить гипотезу

$$H_0 : A_i, i = \ell, \ldots, d \text{ не описывают отличия.}$$

Введем статистику $\Lambda - prime$:

$$\Lambda_\ell^p = \prod_{i=l}^{d} \frac{1}{1+\lambda_i}.$$

Тогда гипотезу выше можно переформулировать так

$$H_0 : \Lambda_\ell^p = 1 \Leftrightarrow \lambda_\ell = \ldots = \lambda_d = 0 \Leftrightarrow rank\mathbf{B} = \ell - 1.$$

Критерий:

$$t = \Lambda_\ell^p \sim \Lambda_{\nu_\mathbf{B}+(\ell-1),\nu_\mathbf{W}-(\ell-1)}.$$

4.9 Последовательный дискриминантный анализ

Опишем подход к feature selection в дискриминантном анализе на обычном языке.

Возникает вопрос отбора признаков. С одной стороны нам бы хотелось, чтобы признаки были независимы, с другой – убрать признаки, которые не влияют на качество разделения. Для начала более формально определим от каких признаков мы хотим избавиться:

- Признаки, которые являются линейной комбинацией других признаков, другими словами имеют большой коэффициент множественной корреляции $\mathbf{R}^2 = \mathbf{R}^2(\xi_i; \{\xi_j \mid j \neq i\})$ (по pooled ковариационной матрице);
- Признаки, которые не влияют на качество разделения. Действуем аналогично пошаговой регрессии. Введем статистику

$$(Partial\Lambda)_i = \Lambda(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p) = \\ = \frac{\Lambda(X_1, \ldots, X_p)}{\Lambda(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)}.$$

  Гипотеза:

  $H_0:$ добавление X_i не влияет на качество разделения $\Leftrightarrow (Partial\Lambda)_i = 1$.

Критерий (принимая во внимание соотношение между распределением $\Lambda_1(\nu_{\mathbf{B}}, \nu_{\mathbf{W}} - p + 1)$ и распределением Фишера):

$$F_i = \frac{1 - (Partial\Lambda)_i/\nu_{\mathbf{B}}}{(Partial\Lambda)_i/(\nu_{\mathbf{W}} - p + 1)} \sim F_{\nu_{\mathbf{B}}}.$$

Далее жадным образом отбираем признаки, влияющие на качество разделения.

4.10 LDA как минимизация эмпирического риска

В данном разделе сведем задачу максимизации апостериорных вероятностей к задаче минимизации эмпирического риска.

Покажем, что задача максимизации эмпирического риска эквивалентна решению обобщенной задачи на собственные числа. Обобщенная задача на собственные вектора имеет вид

$$\mathbf{B}A = \lambda \mathbf{W}A.$$

Заметим, что матрицу $\mathbf{W}$ можно заменить на ковариационную матрицу $\mathbf{\Sigma}$ так как она является суммой $\mathbf{B}$ и $\mathbf{W}$ и такая замена не изменит собственные вектора. В случае двух групп одинакового размера

$$\sum_{i=1}^{2} m(\mu_i - \frac{\mu_1 + \mu_2}{2})(\mu_i - \frac{\mu_1 + \mu_2}{2})^{\mathrm{T}} A = \lambda \mathbf{\Sigma} A.$$

Далее

$$m(\mu_1 - \mu_2)(\mu_{\mathbf{1}} - \mu_2)^{\mathrm{T}} A = \lambda \mathbf{\Sigma} A.$$

В левой части стоит ортогональный проектор на прямую $\langle(\mu_1 - \mu_2)\rangle$. Следовательно, с точностью до константы (которую мы можем «вложить» в собственное число)

$$\lambda \mathbf{\Sigma} A = (\mu_1 - \mu_2), \quad A \propto \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2).$$

Полученный вектор является вектором нормали к разделяющей гиперплоскости и совпадает с тем, что мы получаем в байесовском подходе.

Известно, что LDA в форме обобщенной задачи на собственные вектора эквивалентен каноническому корреляционному анализу [1]. С другой стороны CCA в случае, когда одна из групп признаков одномерная, совпадает с линейной регрессией.

Таким образом, LDA эквивалентно линейной регрессии вида

$$\sum_{i=1}^{N} \left(\mathbf{x}_i^{\mathrm{T}} \beta + \beta_0 - y_i\right)^2 \to \min_{\beta, \beta_0}.$$

[1] См. доказательство например здесь

Как и выше, мы предполагаем, что представителей классов $-1, 1$ поровну, т.е. $\sum_{i=1}^{N} y_i = 0$ и априорные вероятности в LDA равны. Такая задача эквивалентна

$$\sum_{i=1}^{N} \left((\mathbf{x}_i^{\mathrm{T}} \alpha + \alpha_0) y_i - 1 \right)^2 \to \min_{\alpha, \alpha_0},$$

где α и α_0 отличаются от β и β_0 только масштабом. Полученная задача и есть ERM.

Задача максимизации апостериорных вероятностей эквивалентна минимизации аппроксимации эмпирического риска (см. Рис. 1)

$$Q(\alpha, \alpha_0) = (1 - M(\alpha, \alpha_0))^2.$$

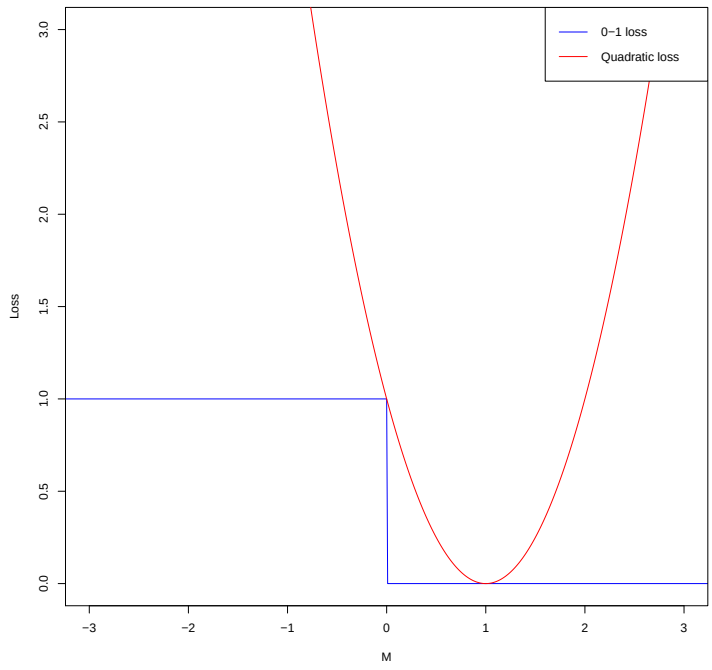

Рис. 1: Аппроксимация функции потерь

5 Логистическая регрессия

Рассмотрим логистическую регрессию как еще один метод построения байесовского классификатора.

Модель задается системой

$$\log \frac{P(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x})}{P(\eta = G_K \mid \boldsymbol{\xi} = \mathbf{x})} = \beta_{i0} + \beta_i^{\mathrm{T}} \mathbf{x}, \quad i = 1, \ldots, K-1.$$

То есть задается $K-1$ log-odds или logit преобразованиями. Заметим, что в знаменателе можно поставить любой класс и оценки вероятностей не поменяются, то есть выбор класса в знаменателе случаен.

Если мы перейдем от логитов к вероятностям, то их сумма будет равна единице

$$P(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x}) = \frac{e^{\beta_{i0} + \beta_i^{\mathrm{T}} \mathbf{x}}}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0} + \beta_k^{\mathrm{T}} \mathbf{x}}}, \quad i = 1, \ldots, K-1,$$
$$P(\eta = G_K \mid \boldsymbol{\xi} = \mathbf{x}) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0} + \beta_k^{\mathrm{T}} \mathbf{x}}}.$$

5.1 Метод максимального правдоподобия и алгоритм Ньютона-Рафсона

Для оценки параметров воспользуемся методом максимального правдоподобия. Рассмотрим логарифм функции максимального правдоподобия

$$\ell(\theta) = \sum_{i=1}^{N} \log P(\eta = G_k \mid \boldsymbol{\xi} = \mathbf{x}_i; \theta), \quad \theta = (\beta_{10}, \beta_1^{\mathrm{T}}, \ldots, \beta_{(K-1)0}, \beta_{K-1}^{\mathrm{T}}).$$

Подробно обсудим случай двух классов $\mathcal{G} = \{0, 1\}$. Обозначим $p(\mathbf{x}, \theta) = P(\eta = 0 \mid \boldsymbol{\xi} = \mathbf{x}; \theta)$ и $1 - p(\mathbf{x}, \theta) = P(\eta = 1 \mid \boldsymbol{\xi} = \mathbf{x}; \theta)$. Тогда логарифм правдоподобия

$$\ell(\beta) = \sum_{i=1}^{N} (y_i \beta^{\mathrm{T}} \mathbf{x}_i - \log(1 + e^{\beta^{\mathrm{T}} \mathbf{x}_i})), \quad \beta = \{\beta_{10}, \beta_1\}.$$

Чтобы максимизировать логарифм правдоподобия, приравниваем производные к нулю, получаем систему из $p+1$ уравнения

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} \mathbf{x}_i (y_i - p(\mathbf{x}_i; \beta)) = 0.$$

Для решения этой системы используем алгоритм Ньютона-Рафсона. Для начала выпишем гессиан логарифма правдоподобия

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} = -\sum_{i=1}^{N} x_i x_i^{\mathrm{T}} p(\mathbf{x}_i; \beta)(1 - p(\mathbf{x}_i; \beta)) = 0.$$

Пусть β^{old} – некоторое начальное приближение вектора коэффициентов β, на каждой итерации он уточняется следующим образом:

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}\right)^{-1} \frac{\partial \ell(\beta)}{\partial\beta},$$

где производные вычисляются в точке β^{old}.

Перейдем к матричным обозначениям. Обозначим $\mathbf{y}$ ответы y_i, $\mathbf{X}$ – матрицу данных, $\mathbf{p} = (p(x_i; \beta^{old}))$, $\mathbf{W}$ – диагональная матрица размером $N \times N$ весов, где iй элемент имеет вид $p(x_i; \beta^{old})(1 - p(x_i; \beta^{old}))$. Тогда

$$\frac{\partial \ell(\beta)}{\partial\beta} = \mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{p})$$
$$\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}} = -\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}.$$

Перепишем шаг алгоритма Ньютона-Рафсона

$$\begin{aligned}\beta^{new} = \beta^{old} + (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{p}) = \\ = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(\mathbf{X}\beta^{old} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})) = \\ = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z}.\end{aligned}$$

Мы переписали итерацию алгоритма как взвешенную регрессию, где в качестве ответа выступает вектор

$$\mathbf{z} = \mathbf{X}\beta^{old} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p}).$$

На каждом шаге $\mathbf{p}$ меняется, а вместе с ним и $\mathbf{W}$, $\mathbf{z}$. Этот алгоритм называется iteratively reweighted least squares (IRLS) так как на каждом шаге решается задача

$$\beta^{new} = \arg\min_{\beta}(\mathbf{z} - \mathbf{X}\beta)^{\mathrm{T}}\mathbf{W}(\mathbf{z} - \mathbf{X}\beta).$$

В качестве начального приближения β^{old} можно взять оценки, полученные с помощью обычной линейной регрессии или просто $\beta^{old} = 0$. Сходимость нам не гарантируется, но обычно алгоритм сходится так как логарифм правдоподобия вогнутый.

5.2 Минимизация эмпирического риска

В логистической регрессии минимизируется аппроксимация:

$$Q(\beta) = \sum_{i=1}^{N} \log(1 + e^{-y_i\beta^{\mathrm{T}}x_i}) \to \min_{\beta},$$

то есть функция потерь имеет вид $\mathcal{L}(M_i(\beta)) = \log(1 + e^{-y_i\beta^{\mathrm{T}}x_i})$ (см. Рис. 2).

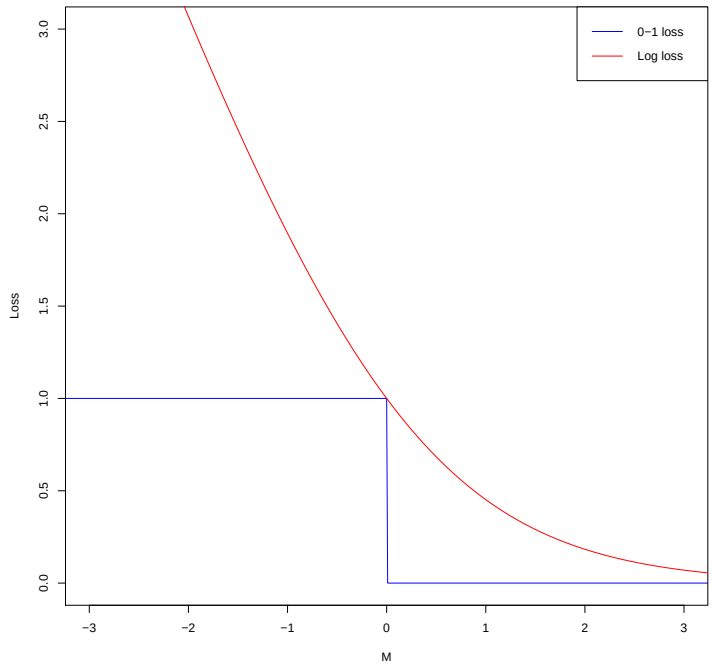

Рис. 2: Аппроксимация функции потерь

5.3 Регуляризация

Аналогично обычной линейной регрессии, можно отбирать признаки (осуществлять feature selection) с помощью LASSO (L1) или аналога Ridge Regression (L2). Для этого максимизируем соответственно

$$\max_{\beta_0,\beta} \sum_{i=1}^{N} \Big(y_i(\beta_0 + \beta^{\mathrm{T}}\mathbf{x}_i) - \log(1 + e^{\beta_0+\beta^{\mathrm{T}}\mathbf{x_i}}) \Big) - \lambda \sum_{j=1}^{p} |\beta_j|,$$

$$\max_{\beta_0,\beta} \sum_{i=1}^{N} \Big(y_i(\beta_0 + \beta^{\mathrm{T}}\mathbf{x}_i) - \log(1 + e^{\beta_0+\beta^{\mathrm{T}}\mathbf{x_i}}) \Big) - \lambda \sum_{j=1}^{p} \beta_j^2.$$

Для нахождения точки максимума можно снова использовать алгоритм Ньютона-Рафсона.

Для feature extraction можно воспользоваться например анализом главных компонент.

6 Логистическая регрессия против линейного дискриминантного анализа

В пункте про линейный дискриминантный анализ мы получили линейную по $\mathbf{x}$ дискриминантную функцию как следствие предположения о нормальном распределении групп и одинаковых ковариационных матрицах. Можно посмотреть на log-posterior odds между классами i и K (или, что то же самое, на разделяющую их гиперплоскость) и получить

$$\log \frac{P(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x})}{P(\eta = G_K \mid \boldsymbol{\xi} = \mathbf{x})} = \\ = -\frac{1}{2}(\mu_i - \mu_K)^{\mathrm{T}}\Sigma^{-1}(\mu_i + \mu_K) + (\mu_i - \mu_K)^{\mathrm{T}}\Sigma^{-1}x + \log(\pi_i/\pi_K) = \\ = \alpha_{i0} + \alpha_i^{\mathrm{T}}\mathbf{x}.$$

С другой стороны, линейная логистическая регрессия имеет линейные логиты по построению

$$\log \frac{P(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x})}{P(\eta = G_K \mid \boldsymbol{\xi} = \mathbf{x})} = \beta_{i0} + \beta_i^{\mathrm{T}}\mathbf{x}.$$

Модели выглядят очень похоже. Различие заключается в том как оцениваются линейные коэффициенты. Логистическая регрессия – более общий подход, мы делаем меньше предположений.

Выпишем совместную плотность X и G

$$P(\boldsymbol{\xi} = \mathbf{x}, \eta = G_i) = P(\mathbf{x})P(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x}).$$

И в линейном дискриминантном анализе, и в логистической регрессии второй множитель выражается как

$$P(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x}) = \frac{e^{\beta_{i0} + \beta_i^{\mathrm{T}}\mathbf{x}}}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0} + \beta_k^{\mathrm{T}}\mathbf{x}}}.$$

В логистической регрессии $P(X)$ – произвольная плотность, а параметры $P(G \mid X)$ оцениваются максимизацией условного правдоподобия (сумму логарифмов условных плотностей классов). Такой подход называют discriminative learning. Решается задача

$$\ell(\theta) = \sum_{i=1}^{N} \log P(\eta = y_i \mid \boldsymbol{\xi} = \mathbf{x}_i; \theta) \to \max_{\theta}.$$

С другой стороны, в LDA мы максимизируем полноценный логарифм функции правдоподобия совместной плотности

$$P(\mathbf{x}, \eta = G_i) = \phi(\mathbf{x}; \mu_i, \Sigma)\pi_i,$$

где $\phi(\mathbf{x}; \mu_i, \Sigma)$ – плотность нормального распределения. Такой подоход называют generative learning. Решается задача

$$\ell(\mu_i, \Sigma) = \sum_{i=1}^{N} \phi(\mathbf{x}_i; \mu_i, \Sigma)\pi_i \to \max_{\mu_i, \Sigma}.$$

Оценив параметры нормального распределения, мы можем подставить их в выражения для логитов. В отличие от логистической регресси, плотность $P(X)$ здесь играет роль. Это смесь распределений

$$P(\mathbf{x}) = \sum_{i=1}^{K} \phi(\mathbf{x}; \mu_i, \Sigma)\pi_i.$$

Возникает вопрос, что нам дает такая модель? Предположение о нормальности распределения дает нам больше информации о параметрах, отсюда меньше дисперсия оценок. С другой стороны, точки, которые находятся далеко от разделяющей плоскости (у которых в логистической регрессии вес будет меньше), влияют на оценку ковариационной матрицы. Это значит, что LDA не является робастным по отношению к выбросам.

В логистической регрессии модель более гибкая, так как у нас меньше ограничений на распределения групп. Отсюда и меньшее количество параметров, которое необходимо оценивать.

Сравнить два подхода можно и с точки зрения минимизации эмпирического риска в терминах ML. Сразу становится ясно, что логистическая регрессия и линейный дискриминантный анализ решают разные задачи так как минимизируют разные аппроксимации эмпирического риска.

7 Непараметрическое оценивание плотностей

Выше мы строили байесовский классификатор исходя из каких-то модельных предположений. Ниже предложен непараметрический способ оценки плотностей.

Локальная непараметрическая оценка Парзена-Розенблата имеет следующий вид:

$$\widehat{p}_h(\mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{p} \frac{1}{h_j} K\left(\frac{z_j - x_{ij}}{h_j}\right),$$

где $K(x)$ – четная и нормированная функция $\int K(x)dx = 1$, которую называют ядром; $h > 0$ – ширина окна, которая выбирается с помощью скользящего контроля (LOO).

Если $K(x)$ – непрерывно, $\int K(x)^2 dx < \infty$ и найдется последовательность h_N такая, что $\lim\limits_{N\to\infty} h_N = 0$ и $\lim\limits_{N\to\infty} Nh_N = \infty$, тогда $\widehat{p}_{h_m}(\mathbf{z}) \to p(x)$ п.в. при $N \to \infty$.

Метод парзеновского окна расширяется на случай произвольной метрики, а также на случай переменной ширины окна. Последнее помогает избежать проблему локальных сгущений в случае сильно неравномерного распределения. Одно и то же значение h приведет к чрезмерному сглаживанию плотности в одних областях пространства и недостаточному в других.

Выбор ядра не влияет на качество оценки, но определяет гладкость функции $\widehat{p}_h$ и влияет на эффективность вычислений.

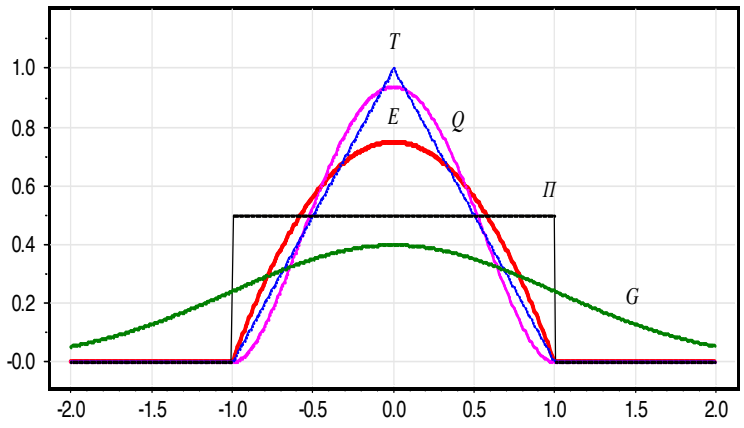

Рис. 3: Различные ядра: E – Епанечикова, Q – Квартическое, T – Треугольное, G – Гауссовское, П – Прямоугольное