

Введение в классификацию

Сергей Николенко

Академия MADE — Mail.Ru

20 февраля 2021 г.

Random facts:

- 20 февраля 395 в Вифлееме был основан первый женский монастырь, а 20 февраля 1819 г. был учреждён Санкт-Петербургский государственный университет
- 20 февраля 1816 г. в Риме прошла премьера оперы «Севильский цирюльник», а 20 февраля 1877 г. Большой театр впервые представил балет «Лебединое озеро»
- 20 февраля 1909 г. на первой странице Le Figaro был опубликован Манифест футуризма; впрочем, Маринетти пришлось заплатить за это как за рекламное объявление
- 20 февраля 1935 г. Каролина Миккельсен стала первой женщиной, побывавшей в Антарктиде
- 20 февраля 1962 г. Джон Гленн трижды облетел Землю, а 20 февраля 1986 г. был выведен на орбиту базовый блок орбитальной станции «Мир»

Введение в классификацию

Задача классификации

- Теперь классификация: определить вектор x в один из K классов C_k .
- В итоге у нас так или иначе всё пространство разобьётся на эти классы.
- Т.е. на самом деле мы ищем *разделяющую поверхность* (decision surface, decision boundary).

Задача классификации

- Как кодировать? Бинарная задача – очень естественно, переменная t , $t = 0$ соответствует \mathcal{C}_1 , $t = 1$ соответствует \mathcal{C}_2 .
- Оценку t можно интерпретировать как вероятность (по крайней мере, мы постараемся, чтобы было можно).
- Если несколько классов – удобно 1-of-K:

$$\mathbf{t} = (0, \dots, 0, 1, 0, \dots)^\top.$$

- Тоже можно интерпретировать как вероятности – или пропорционально им.

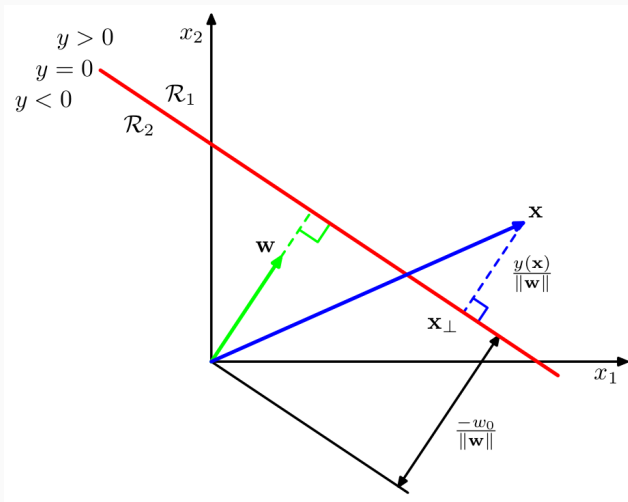
Разделяющая гиперплоскость

- Начнём с геометрии: рассмотрим линейную дискриминантную функцию

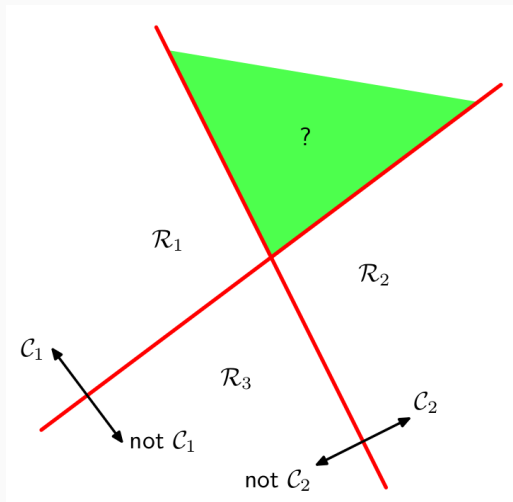
$$y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0.$$

- Это гиперплоскость, и \mathbf{w} – нормаль к ней.
- Расстояние от начала координат до гиперплоскости равно $\frac{-w_0}{\|\mathbf{w}\|}$.
- $y(\mathbf{x})$ связано с расстоянием до гиперплоскости: $d = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$.

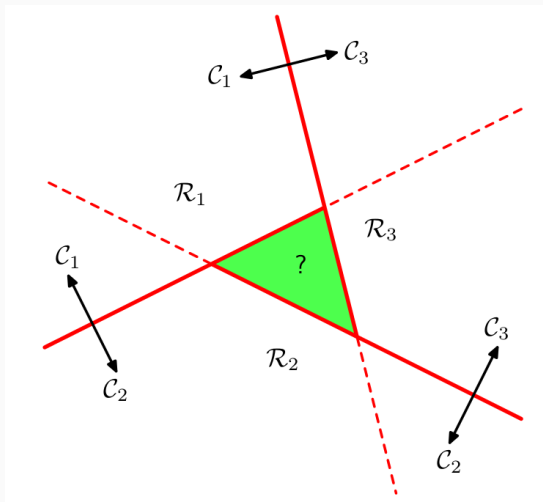
Разделяющая гиперплоскость



- С несколькими классами выходит задача.
- Можно рассмотреть K поверхностей вида «один против всех».
- Можно – $\binom{K}{2}$ поверхностей вида «каждый против каждого».
- Но всё это как-то нехорошо.



Несколько классов

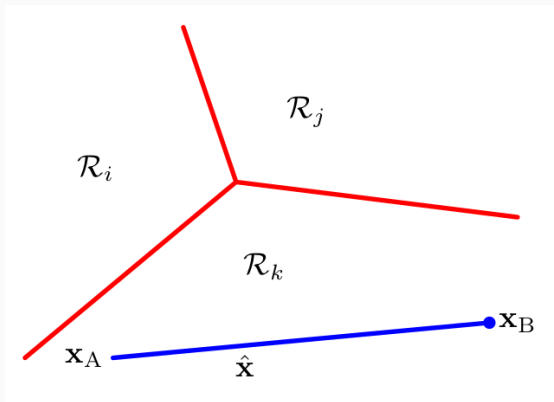


- Лучше рассмотреть единый дискриминант из K линейных функций:

$$y_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0}.$$

- Классифицировать в \mathcal{C}_k , если $y_k(\mathbf{x})$ – максимален.
- Тогда разделяющая поверхность между \mathcal{C}_k и \mathcal{C}_j будет гиперплоскостью вида $y_k(\mathbf{x}) = y_j(\mathbf{x})$:

$$(\mathbf{w}_k - \mathbf{w}_j)^\top \mathbf{x} + (w_{k0} - w_{j0}) = 0.$$



Упражнение. Докажите, что области, соответствующие классам, при таком подходе всегда односвязные и выпуклые.

Метод наименьших квадратов

- Мы снова можем воспользоваться методом наименьших квадратов: запишем $y_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0}$ вместе (спрятав свободный член) как

$$y(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}.$$

- Можно найти \mathbf{W} , оптимизируя сумму квадратов; функция ошибки:

$$E_D(\mathbf{W}) = \frac{1}{2} \text{Tr} \left[(\mathbf{XW} - \mathbf{T})^\top (\mathbf{XW} - \mathbf{T}) \right].$$

- Берём производную, решаем...

- ...получается привычное

$$W = (X^T X)^{-1} X^T T = X^\dagger T,$$

где X^\dagger – псевдообратная Мура-Пенроуза.

- Теперь можно найти и дискриминантную функцию:

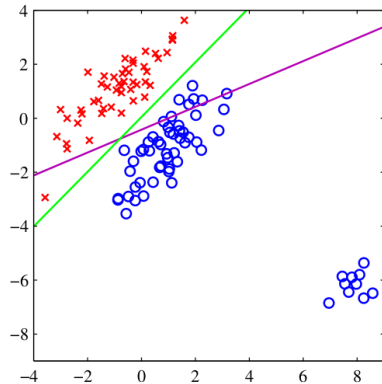
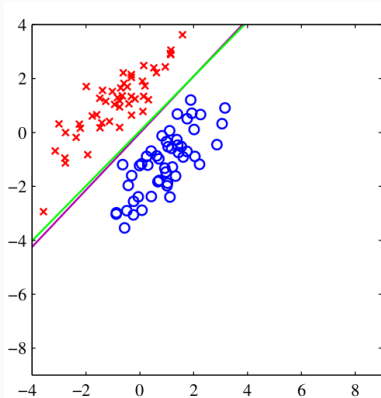
$$y(x) = W^T x = T^T (X^\dagger)^T x.$$

- Это решение сохраняет линейность.

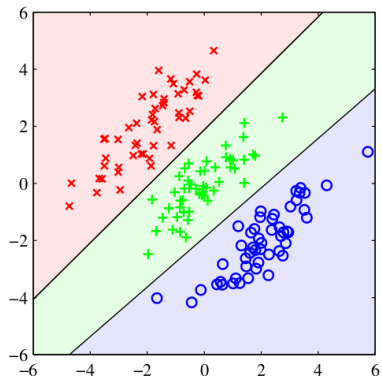
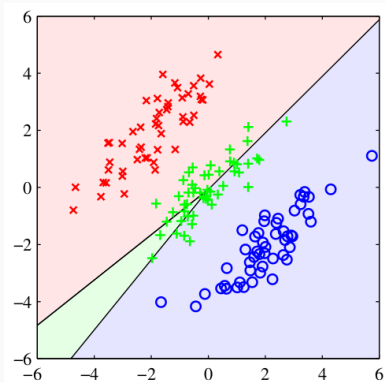
Упражнение. Докажите, что в схеме кодирования 1-of-K предсказания $y_k(\mathbf{x})$ для разных классов при любом \mathbf{x} будут давать в сумме 1. Почему они всё-таки не будут разумными оценками вероятностей?

- Проблемы наименьших квадратов:
 - outliers плохо обрабатываются;
 - «слишком правильные» предсказания добавляют штраф.

Проблемы наименьших квадратов



Проблемы наименьших квадратов



Проблемы наименьших квадратов

- Почему так? Почему наименьшие квадраты так плохо работают?

Проблемы наименьших квадратов

- Почему так? Почему наименьшие квадраты так плохо работают?
- Они предполагают гауссовское распределение ошибки.
- Но, конечно, распределение у бинарных векторов далеко не гауссово.

Линейный дискриминант Фишера

- Другой взгляд на классификацию: в линейном случае мы хотим спроецировать точки в размерность 1 (на нормаль разделяющей гиперплоскости) так, чтобы в этой размерности 1 они хорошо разделялись.
- Т.е. классификация – это такой метод радикального сокращения размерности.
- Давайте посмотрим на классификацию с этих позиций и попробуем добиться оптимальности в каком-то смысле.

Линейный дискриминант Фишера

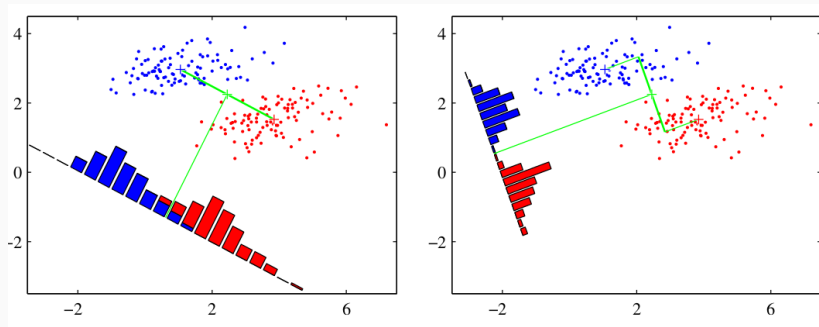
- Рассмотрим два класса \mathcal{C}_1 и \mathcal{C}_2 с N_1 и N_2 точками.
- Первая идея – надо найти серединный перпендикуляр между центрами кластеров

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{\mathcal{C}_1} \mathbf{x}, \text{ и } \mathbf{m}_2 = \frac{1}{N_2} \sum_{\mathcal{C}_2} \mathbf{x},$$

т.е. максимизировать $\mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1)$.

- Надо ещё добавить ограничение $\|\mathbf{w}\| = 1$, но всё равно не ахти как работает.

Линейный дискриминант Фишера



Чем левая картинка хуже правой?

Линейный дискриминант Фишера

- Слева больше дисперсия каждого кластера.
- Идея: минимизировать перекрытие классов, оптимизируя и проекцию расстояния, и дисперсию.
- Выборочные дисперсии в проекции: для $y_n = \mathbf{w}^T \mathbf{x}_n$

$$s_1 = \sum_{n \in \mathcal{C}_1} (y_n - m_1)^2 \text{ и } s_2 = \sum_{n \in \mathcal{C}_2} (y_n - m_2)^2.$$

Линейный дискриминант Фишера

- Критерий Фишера:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}}, \text{ где}$$

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top,$$

$$\mathbf{S}_W = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top.$$

(between-class covariance и within-class covariance).

- Дифференцируя по \mathbf{w} ...

Линейный дискриминант Фишера

- ...получим, что $J(\mathbf{w})$ максимален при

$$(\mathbf{w}^\top \mathbf{S}_B \mathbf{w}) \mathbf{S}_W \mathbf{w} = (\mathbf{w}^\top \mathbf{S}_W \mathbf{w}) \mathbf{S}_B \mathbf{w}.$$

- Т.к. $\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top$, $\mathbf{S}_B \mathbf{w}$ всё равно будет в направлении $\mathbf{m}_2 - \mathbf{m}_1$, а длина \mathbf{w} нас не интересует.
- Поэтому получается

$$\mathbf{w} \propto \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1).$$

- В итоге мы выбрали направление проекции, и осталось только разделить данные на этой проекции.

Линейный дискриминант Фишера

- Любопытно, что дискриминант Фишера тоже можно получить из наименьших квадратов.
- Давайте для класса C_1 выберем целевое значение $\frac{N_1+N_2}{N_1}$, а для класса C_2 возьмём $-\frac{N_1+N_2}{N_2}$.

Упражнение. Докажите, что при таких целевых значениях наименьшие квадраты – это дискриминант Фишера.

Линейный дискриминант Фишера

- А что будет с несколькими классами? Рассмотрим $\mathbf{y} = \mathbf{W}^\top \mathbf{x}$, обобщим внутреннюю дисперсию как

$$\mathbf{S}_W = \sum_{k=1}^K \mathbf{S}_k = \sum_{k=1}^K \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k) (\mathbf{x}_n - \mathbf{m}_k)^\top.$$

- Чтобы обобщить внешнюю (межклассовую) дисперсию, просто возьмём остаток полной дисперсии

$$\mathbf{S}_T = \sum_n (\mathbf{x}_n - \mathbf{m}) (\mathbf{x}_n - \mathbf{m})^\top,$$

$$\mathbf{S}_B = \mathbf{S}_T - \mathbf{S}_W.$$

Линейный дискриминант Фишера

- Обобщить критерий можно разными способами, например:

$$J(W) = \text{Tr} [s_W^{-1} s_B] ,$$

где s – ковариации в пространстве проекций на y :

$$s_W = \sum_{k=1}^K \sum_{n \in \mathcal{C}_k} (y_n - \mu_k) (y_n - \mu_k)^\top ,$$

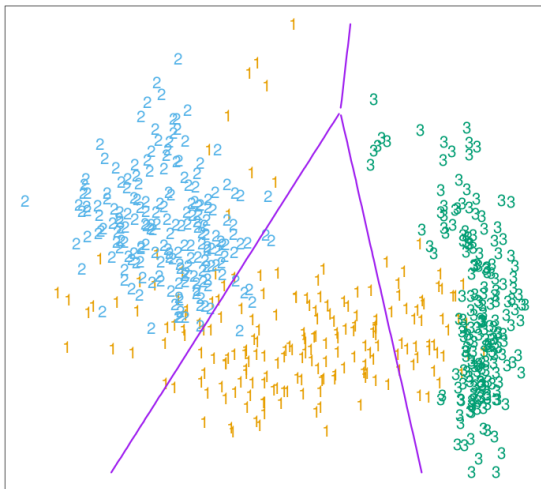
$$s_B = \sum_{k=1}^K N_k (\mu_k - \mu) (\mu_k - \mu)^\top ,$$

где $\mu_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} y_n$.

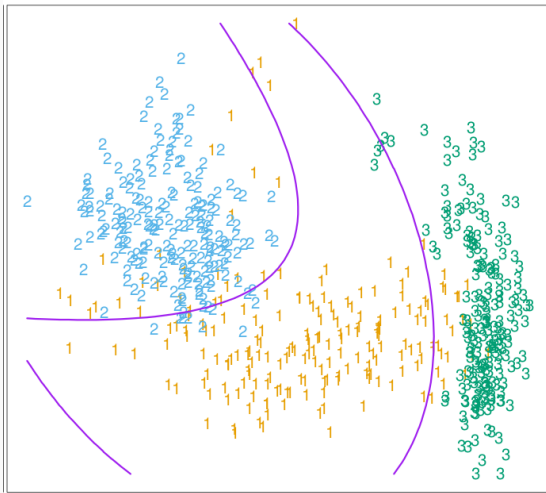
LDA и QDA

- Мы учились проводить разделяющие гиперплоскости.
- Но как же нелинейные поверхности?
- Можно делать нелинейные из линейных, увеличивая размерность.

Нелинейные поверхности



Нелинейные поверхности



- Теперь классификация через генеративные модели: давайте каждому классу сопоставим плотность $p(\mathbf{x} | \mathcal{C}_k)$, найдём априорные распределения $p(\mathcal{C}_k)$, будем искать $p(\mathcal{C}_k | \mathbf{x})$ по теореме Байеса.
- Для двух классов:

$$p(\mathcal{C}_1 | \mathbf{x}) = \frac{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x} | \mathcal{C}_2)p(\mathcal{C}_2)}.$$

- Перепишем:

$$p(\mathcal{C}_1 | \mathbf{x}) = \frac{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x} | \mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1 + e^{-a}} = \sigma(a),$$

где

$$a = \ln \frac{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} | \mathcal{C}_2)p(\mathcal{C}_2)}, \quad \sigma(a) = \frac{1}{1 + e^{-a}}.$$

- $\sigma(a)$ – логистический сигмоид:

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

- $\sigma(-a) = 1 - \sigma(a)$.
- $a = \ln\left(\frac{\sigma}{1-\sigma}\right)$ – логит-функция.

Упражнение. Докажите эти свойства.

- В случае нескольких классов получится

$$p(\mathcal{C}_k | \mathbf{x}) = \frac{p(\mathbf{x} | \mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x} | \mathcal{C}_j)p(\mathcal{C}_j)} = \frac{e^{a_k}}{\sum_j e^{a_j}}.$$

- Здесь $a_k = \ln p(\mathbf{x} | \mathcal{C}_k)p(\mathcal{C}_k)$.
- $\frac{e^{a_k}}{\sum_j e^{a_j}}$ – нормализованная экспонента, или softmax-функция (сглаженный максимум).

- Давайте рассмотрим гауссовы распределения для классов:

$$p(\mathbf{x} \mid \mathcal{C}_k) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma).$$

- Сначала пусть Σ у всех одинаковые, а классов всего два.
- Посчитаем логистический сигмоид...

- ...получится

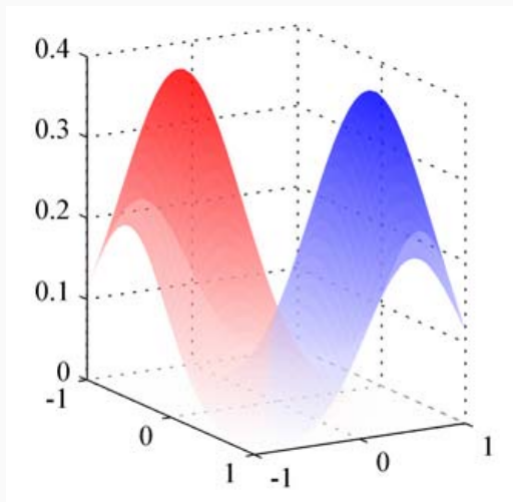
$$p(C_1 | \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + w_0), \text{ где}$$

$$\mathbf{w} = \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

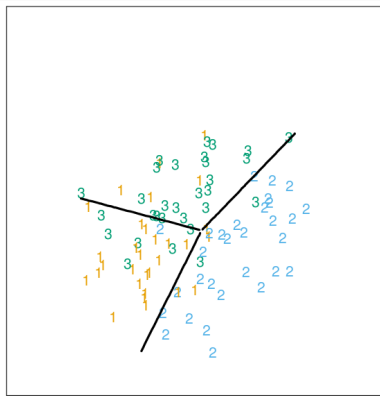
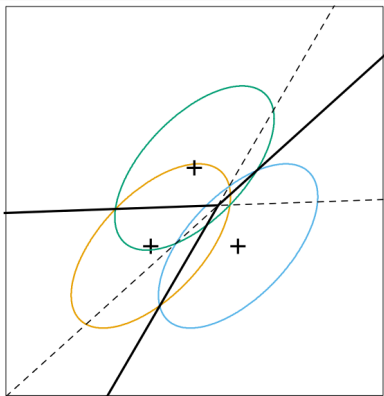
$$w_0 = -\frac{1}{2} \boldsymbol{\mu}_1^\top \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^\top \Sigma^{-1} \boldsymbol{\mu}_2 + \ln \frac{p(C_1)}{p(C_2)}.$$

- Т.е. в аргументе сигмоида получается линейная функция от \mathbf{x} .
Поверхности уровня – это когда $p(C_1 | \mathbf{x})$ постоянно, т.е.
гиперплоскости в пространстве \mathbf{x} . Априорные вероятности
 $p(C_k)$ просто сдвигают эти гиперплоскости.

Разделяющая гиперплоскость

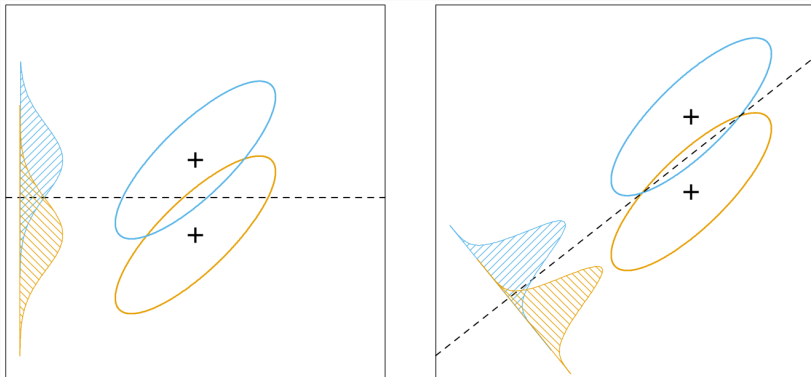


Разделяющая гиперплоскость



Дискриминант Фишера

Кстати, с дискриминантом Фишера эта разделяющая поверхность отлично сходится.



- С несколькими классами получится тоже примерно так же:

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \Sigma^{-1} \boldsymbol{\mu}_k + \ln \pi_k,$$

где $\pi_k = p(C_k)$.

- Получились линейные $\delta_k(\mathbf{x})$, и опять разделяющие поверхности линейные (тут разделяющие поверхности – когда две максимальных вероятности равны).
- Этот метод называется LDA – linear discriminant analysis.

Метод максимального правдоподобия

- Как оценить распределения $p(\mathbf{x} \mid \mathcal{C}_k)$, если даны только данные?
- Можно по методу максимального правдоподобия.
- Опять рассмотрим тот же пример: два класса, гауссианы с одинаковой матрицей ковариаций, и есть $D = \{\mathbf{x}_n, t_n\}_{n=1}^N$, где $t_n = 1$ значит \mathcal{C}_1 , $t_n = 0$ значит \mathcal{C}_2 .
- Обозначим $p(\mathcal{C}_1) = \pi$, $p(\mathcal{C}_2) = 1 - \pi$.

Метод максимального правдоподобия

- Для одной точки в классе \mathcal{C}_1 :

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n | \mathcal{C}_1) = \pi \mathcal{N}(\mathbf{x}_n | \mu_1, \Sigma).$$

- В классе \mathcal{C}_2 :

$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathbf{x}_n | \mathcal{C}_2) = (1 - \pi) \mathcal{N}(\mathbf{x}_n | \mu_2, \Sigma).$$

- Функция правдоподобия:

$$\begin{aligned} p(\mathbf{t} | \pi, \mu_1, \mu_2, \Sigma) = \\ = \prod_{n=1}^N [\pi \mathcal{N}(\mathbf{x}_n | \mu_1, \Sigma)]^{t_n} [(1 - \pi) \mathcal{N}(\mathbf{x}_n | \mu_2, \Sigma)]^{1-t_n}. \end{aligned}$$

Метод максимального правдоподобия

- Максимизируем логарифм правдоподобия. Сначала по π , там останется только

$$\sum_{n=1}^N [t_n \ln \pi + (1 - t_n) \ln(1 - \pi)],$$

и, взяв производную, получим, совершенно неожиданно,

$$\hat{\pi} = \frac{N_1}{N_1 + N_2}.$$

Метод максимального правдоподобия

- Теперь по μ_1 ; всё, что зависит от μ_1 :

$$\sum_n t_n \ln \mathcal{N}(\mathbf{x}_n \mid \mu_1, \Sigma) = -\frac{1}{2} \sum_n t_n (\mathbf{x}_n - \mu_1)^\top \Sigma^{-1} (\mathbf{x}_n - \mu_1) + C.$$

- Берём производную, и получается, опять внезапно,

$$\hat{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^N t_n \mathbf{x}_n.$$

- Аналогично,

$$\hat{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^N (1 - t_n) \mathbf{x}_n.$$

Метод максимального правдоподобия

- Для матрицы ковариаций придётся постараться; в результате получится

$$\hat{\Sigma} = \frac{N_1}{N_1 + N_2} \mathbf{S}_1 + \frac{N_2}{N_1 + N_2} \mathbf{S}_2, \text{ где}$$
$$\mathbf{S}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \boldsymbol{\mu}_1) (\mathbf{x}_n - \boldsymbol{\mu}_1)^\top,$$
$$\mathbf{S}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \boldsymbol{\mu}_2) (\mathbf{x}_n - \boldsymbol{\mu}_2)^\top.$$

- Тоже совершенно неожиданно: взвешенное среднее оценок для двух матриц ковариаций.

- Это самым прямым образом обобщается на случай нескольких классов.

Упражнение. Сделайте это.

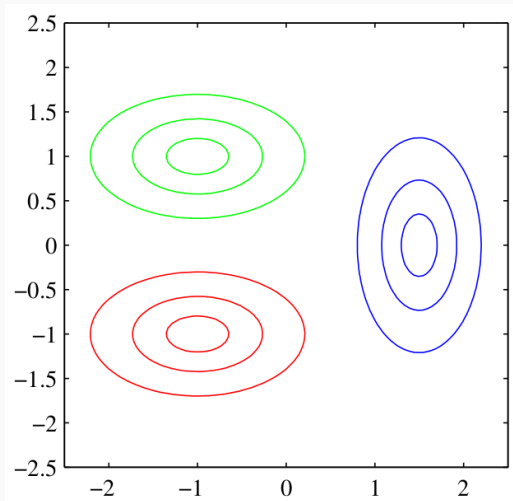
- А вот с разными матрицами ковариаций уже будет по-другому.
- Квадратичные члены не сократятся.
- Разделяющие поверхности станут квадратичными; QDA – quadratic discriminant analysis.

- В QDA получится

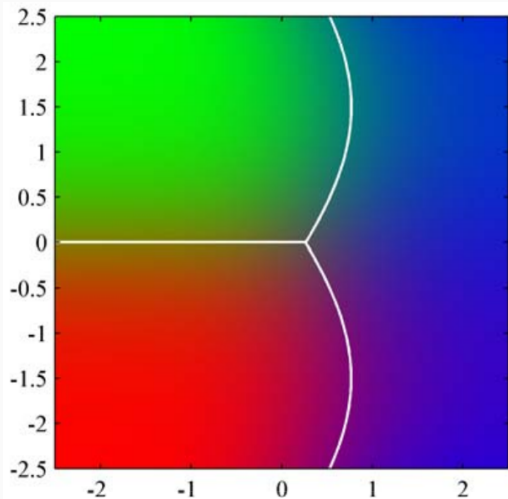
$$\delta_k(\mathbf{x}) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) + \log \pi_k.$$

- Разделяющая поверхность между \mathcal{C}_i и \mathcal{C}_j – это $\{\mathbf{x} \mid \delta_i(\mathbf{x}) = \delta_j(\mathbf{x})\}$.
- Оценки максимального правдоподобия такие же, только надо отдельно матрицы ковариаций оценивать.

Разные матрицы ковариаций

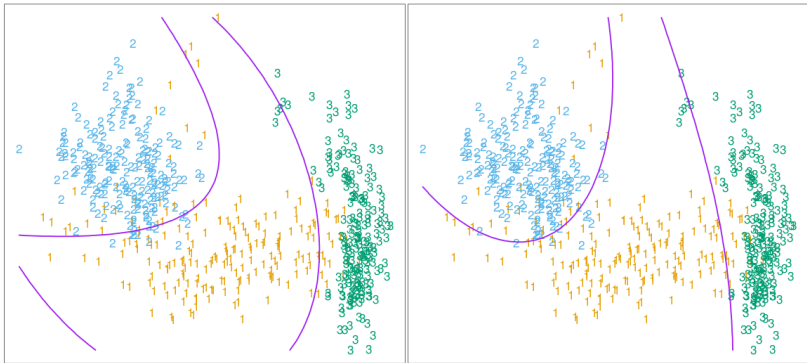


Разные матрицы ковариаций



LDA vs. QDA

Разница между LDA с квадратичными членами и QDA обычно невелика.



- LDA и QDA неплохо работают на практике. Часто это первая идея в классификации.
- Число параметров:
 - у LDA $(K - 1)(d + 1)$ параметр: по $d + 1$ на каждую разницу вида $\delta_k(\mathbf{x}) - \delta_K(\mathbf{x})$;
 - у QDA $(K - 1)(d(d + 3)/2 + 1)$ параметр, но он выглядит гораздо лучше своих лет.

- Почему хорошо работают?
- Скорее всего, потому, что линейные и квадратичные оценки достаточно стабильны: даже если *bias* относительно большой (как будет, если данные всё-таки не гауссианами порождены), *variance* будет маленькой.

- Компромисс между LDA и QDA – регуляризованный дискриминантный анализ, RDA.
- Стынем ковариации каждого класса к общей матрице ковариаций:

$$\hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1 - \alpha) \hat{\Sigma},$$

где $\hat{\Sigma}_k$ – оценка из QDA, $\hat{\Sigma}$ – оценка из LDA.

- Или стынем к единичной матрице:

$$\hat{\Sigma}_k(\gamma) = \gamma \hat{\Sigma}_k + (1 - \gamma) \hat{\sigma}^2 I.$$

- Предположим, что размерность d больше, чем число классов K .
- Тогда центроиды классов $\hat{\mu}_k$ лежат в подпространстве размерности $\leq K - 1$.
- И когда мы определяем ближайший центроид, нам достаточно считать расстояния только в этом подпространстве.
- Таким образом, можно сократить ранг задачи.

- Куда именно проецировать? Не обязательно само подпространство, порождённое центроидами, будет оптимальным.
- Это мы уже проходили: для размерности 1 это линейный дискриминант Фишера.
- Это он и есть: оптимальное подпространство будет там, где межклассовая дисперсия максимальна по отношению к внутриклассовой.

Логистическая регрессия

- Итак, мы рассмотрели логистический сигмоид:

$$p(C_1 | \mathbf{x}) = \frac{p(\mathbf{x} | C_1)p(C_1)}{p(\mathbf{x} | C_1)p(C_1) + p(\mathbf{x} | C_2)p(C_2)} = \frac{1}{1 + e^{-a}} = \sigma(a),$$

$$\text{где } a = \ln \frac{p(\mathbf{x} | C_1)p(C_1)}{p(\mathbf{x} | C_2)p(C_2)}, \quad \sigma(a) = \frac{1}{1 + e^{-a}}.$$

- Вывели из него LDA и QDA, обучили их методом максимального правдоподобия, а потом отвлеклись на naïve Bayes.

- Возвращаемся к задаче классификации.
- Два класса, и апостериорное распределение – логистический сигмоид на линейной функции:

$$p(C_1 | \phi) = y(\phi) = \sigma(\mathbf{w}^\top \phi), \quad p(C_2 | \phi) = 1 - p(C_1 | \phi).$$

- *Логистическая регрессия* – это когда мы напрямую оптимизируем \mathbf{w} .

- Для датасета $\{\phi_n, t_n\}$, $t_n \in \{0, 1\}$, $\phi_n = \phi(\mathbf{x}_n)$:

$$p(\mathbf{t} \mid \mathbf{w}) = \prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n}, \quad y_n = p(\mathcal{C}_1 \mid \phi_n).$$

- Ищем параметры максимального правдоподобия, минимизируя $-\ln p(\mathbf{t} \mid \mathbf{w})$:

$$E(\mathbf{w}) = -\ln p(\mathbf{t} \mid \mathbf{w}) = -\sum_{n=1}^N [t_n \ln y_n + (1 - t_n) \ln(1 - y_n)].$$

- Пользуясь тем, что $\sigma' = \sigma(1 - \sigma)$, берём градиент (похоже на перцептрон):

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \phi_n.$$

- Если теперь сделать градиентный спуск, получим как раз разделяющую поверхность.
- Заметим, правда, что если данные действительно разделимы, то может получиться жуткий оверфиттинг: $\|\mathbf{w}\| \rightarrow \infty$, и сигмоид превращается в функцию Хевисайда. Надо регуляризовать.

- В логистической регрессии не получается замкнутого решения из-за сигмоида.
- Но функция $E(\mathbf{w})$ всё равно выпуклая, и можно воспользоваться методом Ньютона-Рапсона – на каждом шаге использовать локальную квадратичную аппроксимацию к функции ошибки:

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \mathbf{H}^{-1} \nabla E(\mathbf{w}),$$

где \mathbf{H} (Hessian) – матрица вторых производных $E(\mathbf{w})$.

- Замечание: давайте применим Ньютона-Рапсона к обычной линейной регрессии с квадратической ошибкой:

$$\begin{aligned}\nabla E(\mathbf{w}) &= \sum_{n=1}^N (\mathbf{w}^\top \phi_n - t_n) \phi_n = \mathbf{\Phi}^\top \mathbf{\Phi} \mathbf{w} - \mathbf{\Phi}^\top \mathbf{t}, \\ \nabla \nabla E(\mathbf{w}) &= \sum_{n=1}^N \phi_n \phi_n^\top = \mathbf{\Phi}^\top \mathbf{\Phi},\end{aligned}$$

и шаг оптимизации будет

$$\begin{aligned}\mathbf{w}^{\text{new}} &= \mathbf{w}^{\text{old}} - \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \left[\mathbf{\Phi}^\top \mathbf{\Phi} \mathbf{w}^{\text{old}} - \mathbf{\Phi}^\top \mathbf{t}\right] = \\ &= \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \mathbf{t},\end{aligned}$$

т.е. мы за один шаг придём к решению.

- Для логистической регрессии:

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \phi_n = \mathbf{\Phi}^T (\mathbf{y} - \mathbf{t}),$$

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N y_n (1 - y_n) \phi_n \phi_n^T = \mathbf{\Phi}^T \mathbf{R} \mathbf{\Phi}$$

для диагональной матрицы \mathbf{R} с $R_{nn} = y_n(1 - y_n)$.

- Формула шага оптимизации:

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \left(\Phi^{\top} R \Phi \right)^{-1} \Phi^{\top} (\mathbf{y} - \mathbf{t}) = \left(\Phi^{\top} R \Phi \right)^{-1} \Phi^{\top} R \mathbf{z},$$

где $\mathbf{z} = \Phi \mathbf{w}^{\text{old}} - R^{-1} (\mathbf{y} - \mathbf{t})$.

- Получилось как бы решение взвешенной задачи минимизации квадратического отклонения с матрицей весов R .
- Отсюда название: iterative reweighted least squares (IRLS).

- В случае нескольких классов

$$p(\mathcal{C}_k \mid \phi) = y_k(\phi) = \frac{e^{a_k}}{\sum_j e^{a_j}} \text{ для } a_k = \mathbf{w}_k^\top \phi.$$

- Опять выпишем максимальное правдоподобие; во-первых,

$$\frac{\partial y_k}{\partial a_j} = y_k ([k = j] - y_j) .$$

- Теперь запишем правдоподобие – для схемы кодирования 1-of-K будет целевой вектор \mathbf{t}_n и правдоподобие

$$p(\mathbf{T} \mid \mathbf{w}_1, \dots, \mathbf{w}_K) = \prod_{n=1}^N \prod_{k=1}^K p(c_k \mid \phi_n)^{t_{nk}} = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}$$

для $y_{nk} = y_k(\phi_n)$; берём логарифм:

$$E(\mathbf{w}_1, \dots, \mathbf{w}_K) = -\ln p(\mathbf{T} \mid \mathbf{w}_1, \dots, \mathbf{w}_K) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}, \text{ и}$$

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) = -\sum_{n=1}^N (y_{nj} - t_{nj}) \phi_n.$$

- Оптимизировать опять можно по Ньютону-Рапсону; гессиан получится как

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) = - \sum_{n=1}^N y_{nk} ([k=j] - y_{nj}) \phi_n \phi_n^T.$$

- А что если у нас другая форма сигмоида?
- Мы по-прежнему в той же постановке: два класса, $p(t = 1 | a) = f(a)$, $a = \mathbf{w}^\top \phi$, f – функция активации.
- Давайте установим функцию активации с порогом θ : для каждого ϕ_n , вычисляем $a_n = \mathbf{w}^\top \phi_n$, и

$$\begin{cases} t_n = 1, & \text{если } a_n \geq \theta, \\ t_n = 0, & \text{если } a_n < \theta. \end{cases}$$

- Если θ берётся по распределению $p(\theta)$, это соответствует

$$f(a) = \int_{-\infty}^a p(\theta) d\theta.$$

- Пусть, например, $p(\theta)$ – гауссиан с нулевым средним и единичной дисперсией. Тогда

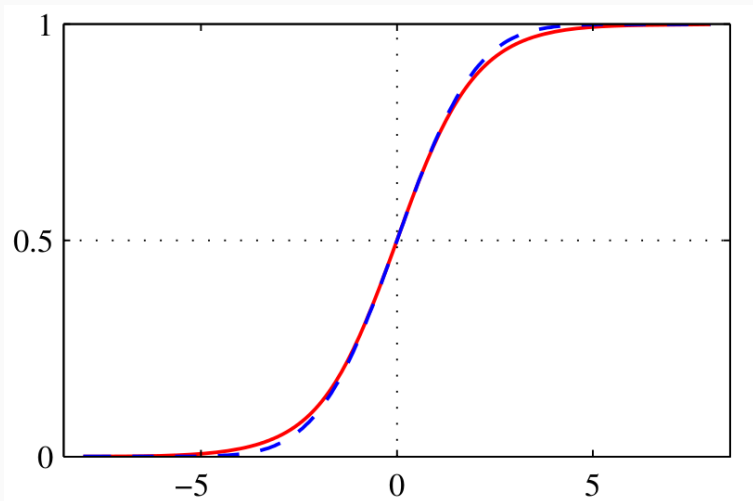
$$f(a) = \Phi(a) = \int_{-\infty}^a \mathcal{N}(\theta \mid 0, 1) d\theta.$$

- Это называется *пробит-функцией* (probit); неэлементарная, но тесно связана с

$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-\frac{\theta^2}{2}} d\theta :$$

$$\Phi(a) = \frac{1}{2} \left[1 + \frac{1}{\sqrt{2}} \operatorname{erf}(a) \right] .$$

- Пробит-регрессия – это модель с пробит-функцией активации.



Логистическая регрессия по-байесовски

- Теперь давайте обработаем логистическую регрессию по-байесовски.
- Логистическую регрессию так просто не выпишешь, как линейную – точного ответа из произведения логистических сигмоидов не получается.
- Будем приближать по Лапласу.

Байесовская логистическая регрессия

- Априорное распределение выберем гауссовским:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

- Тогда апостериорное будет

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{t}) &\propto p(\mathbf{w})p(\mathbf{t} \mid \mathbf{w}), \text{ и} \\ \ln p(\mathbf{w} \mid \mathbf{t}) &= -\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\mathbf{w} - \boldsymbol{\mu}_0) \\ &\quad + \sum_{n=1}^N [t_n \ln y_n + (1 - t_n) \ln(1 - y_n)] + \text{const}, \\ \text{где } y_n &= \sigma(\mathbf{w}^\top \boldsymbol{\phi}_n). \end{aligned}$$

- Чтобы приблизить, сначала находим максимум \mathbf{w}_{MAP} , а потом матрица ковариаций – это матрица вторых производных

$$\mathbf{\Sigma}_N = -\nabla \nabla \ln p(\mathbf{w} \mid \mathbf{t}) = \mathbf{\Sigma}_0^{-1} + \sum_{n=1}^N y_n(1 - y_n) \phi_n \phi_n^{\top}.$$

- Наше приближение – это

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{w}_{\text{MAP}}, \mathbf{\Sigma}_N).$$

- Теперь можно описать байесовское предсказание:

$$p(C_1 | \phi, \mathbf{t}) = \int p(C_1 | \phi, \mathbf{w})p(\mathbf{w} | \mathbf{t})d\mathbf{w} \approx \int \sigma(\mathbf{w}^\top \phi)q(\mathbf{w})d\mathbf{w}.$$

- Заметим, что $\sigma(\mathbf{w}^\top \phi)$ зависит от \mathbf{w} только через его проекцию на ϕ .
- Обозначим $a = \mathbf{w}^\top \phi$:

$$\sigma(\mathbf{w}^\top \phi) = \int \delta(a - \mathbf{w}^\top \phi)\sigma(a)da.$$

- $\sigma(\mathbf{w}^\top \phi) = \int \delta(a - \mathbf{w}^\top \phi) \sigma(a) da$, а значит,

$$\int \sigma(\mathbf{w}^\top \phi) q(\mathbf{w}) d\mathbf{w} = \int \sigma(a) p(a) da,$$

$$\text{где } p(a) = \int \delta(a - \mathbf{w}^\top \phi) q(\mathbf{w}) d\mathbf{w}.$$

- $p(a)$ – это маргинализация гауссиана $q(\mathbf{w})$, где мы интегрируем по всему, что ортогонально ϕ .

Байесовская логистическая регрессия

- $p(a)$ – это маргинализация гауссиана $q(\mathbf{w})$, где мы интегрируем по всему, что ортогонально ϕ .
- Значит, $p(a)$ – тоже гауссиан; найдём его моменты:

$$\begin{aligned}\mu_a &= \mathbb{E}[a] = \int a p(a) da = \int q(\mathbf{w}) \mathbf{w}^\top \phi d\mathbf{w} = \mathbf{w}_{\text{MAP}}^\top \phi, \\ \sigma_a^2 &= \int (a^2 - \mathbb{E}[a])^2 p(a) da = \\ &= \int q(\mathbf{w}) [(\mathbf{w}^\top \phi)^2 - (\mu_N^\top \phi)^2]^2 d\mathbf{w} = \phi^\top \Sigma_N \phi.\end{aligned}$$

- Итого получили, что

$$p(\mathcal{C}_1 | \mathbf{t}) = \int \sigma(a) p(a) da = \int \sigma(a) \mathcal{N}(a | \mu_a, \sigma_a^2) da.$$

- $p(C_1 | \mathbf{t}) = \int \sigma(a) \mathcal{N}(a | \mu_a, \sigma_a^2) da.$
- Этот интеграл так просто не взять, потому что сигмоид сложный, но можно приблизить, если приблизить $\sigma(a)$ через пробит: $\sigma(a) \approx \Phi(\lambda a)$ для $\lambda = \sqrt{\pi/8}$.

Упражнение. Докажите, что для $\lambda = \sqrt{\pi/8}$ у σ и Φ одинаковый наклон в нуле.

- А если мы перейдём к пробит-функции, то её свёртка с гауссианом будет просто другим пробитом:

$$\int \Phi(\lambda a) \mathcal{N}(a \mid \mu, \sigma^2) da = \Phi\left(\frac{\mu}{\sqrt{\frac{1}{\lambda^2} + \sigma^2}}\right).$$

Упражнение. Докажите это.

- В итоге получается аппроксимация

$$\int \sigma(a) \mathcal{N}(a \mid \mu, \sigma^2) da \approx \sigma(\kappa(\sigma^2)\mu),$$

$$\text{где } \kappa(\sigma^2) = \frac{1}{\sqrt{1 + \frac{\pi}{8}\sigma^2}}.$$

Байесовская логистическая регрессия

- И теперь, собирая всё вместе, мы получили распределение предсказаний:

$$p(\mathcal{C}_1 | \phi, \mathbf{t}) = \sigma(\kappa(\sigma_a^2)\mu_a), \text{ где}$$

$$\mu_a = \mathbf{w}_{\text{MAP}}^\top \phi,$$

$$\sigma_a^2 = \phi^\top \Sigma_N \phi,$$

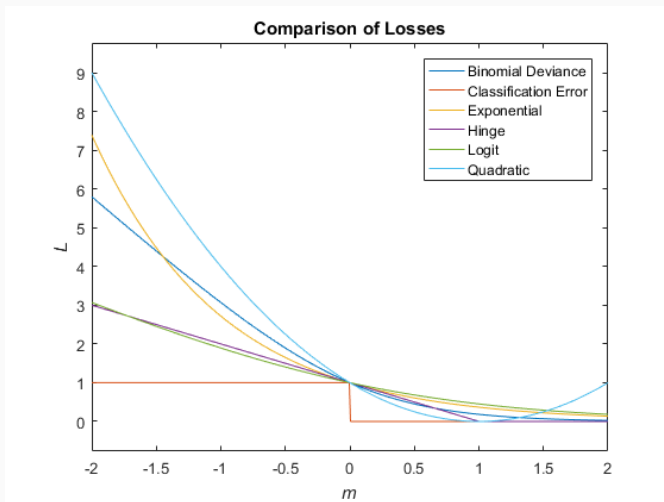
$$\kappa(\sigma^2) = \frac{1}{\sqrt{1 + \frac{\pi}{8}\sigma^2}}.$$

- Кстати, разделяющая поверхность $p(\mathcal{C}_1 | \phi, \mathbf{t}) = \frac{1}{2}$ задаётся уравнением $\mu_a = 0$, и тут нет никакой разницы с просто использованием \mathbf{w}_{MAP} . Разница будет только для более сложных критериев.

Функции потерь в классификации

- И напоследок немножко другой взгляд: разные методы классификации отличаются друг от друга тем, какую функцию ошибки они оптимизируют.
- У классификации проблема с «правильной» функцией ошибки, то есть ошибкой собственно классификации:
 - она и не везде дифференцируема,
 - и производная её никому не нужна.
- Давайте посмотрим на разные функции потерь (loss functions); мы уже несколько видели, но ещё немало осталось.

Функции потерь в классификации



Спасибо!

Спасибо за внимание!