Исследование эффективности некоторых линейных методов классификации на модельных распределениях

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

«Интеллектуализация обработки информации»
(ИОИ-11), Испания, г. Барселона, 10–14 октября 2016 г.

Проблема сравнения методов построения решающих функций

Какие методы дают лучшие решения на задачах www.kaggle.com:

- 50% задач — градиентный бустинг на деревьях (XGBoost).
- 50% задач — градиентный бустинг на деревьях в комбинации с SVM, нейронными сетями, логистической регрессией, k-NN.

Особенность ресурса — сильная заинтересованность участников в качестве решений, непредвзятость в выборе методов.

Цель работы

Рассматривается проблема построения вероятностных моделей, позволяющих выявлять свойства методов построения решающих функций и проводить исследование этих методов.

В частности, ставилась задача построения моделей, на которых заданный метод наиболее эффективен среди сравниваемых методов, а также задача построения моделей максимального правдоподобия.

Практическая цель — понять, за счёт каких свойств реальных задач (свойств даных) один метод классификации превосходит по качеству решений другой метод.

Подходы к построению вероятностных моделей

- Toy examples.
- Приближение к реальным данным в статистическом смысле.
- Построение «эталонных» вероятностных моделей для исследования и сравнения методов построения решающих функций.

Под эталонной моделью понимается вероятностная модель, на которой наиболее выраженно проявляется некоторое свойство исследуемого метода, например, модель, на которой метод демонстрирует наибольшее превосходство, или модель, на которой проявляется некоторый недостаток метода (например, неустойчивость к «выбросам»).

Система «Полигон» — 1980-е

Лбов Г.С., Старцева Н.Г. Сравнение алгоритмов распознавания с помощью программной системы «Полигон» // Анализ данных и знаний в экспертных системах. Новосибирск, 1990. Вып. 134: Вычислительные системы. С. 56–66.

Принципы:

- для каждого метода включается «эталонная» задача,
- на «своей» задаче метод должен работать лучше других,
- возможно оценить степень универсальности метода,
- тестовая единица - таблица данных.

Основные понятия

Пусть X – пространство значений переменных, используемых для прогноза,
$Y = \{0, 1\}$ – пространство значений прогнозируемых переменных,
C – множество всех вероятностных мер на заданной σ-алгебре подмножеств множества $D = X \times Y$.

При каждом $c \in C$ имеем вероятностное пространство: $\langle D, B, \mathsf{P}_c \rangle$, где B – σ-алгебра, P_c – вероятностная мера.

Метод построения решающих функций

Решающей функцией (алгоритмом классификации) называется соответствие $\lambda\colon X \to Y$.
Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}\colon Y^2 \to [0, \infty)$.

Под риском будем понимать средние потери:

$$R(c, \lambda) = \mathsf{E}\mathcal{L}(y, \lambda(x)) = \int_D \mathcal{L}(y, \lambda(x))\ \mathsf{P}_c(dx, dy),$$

$x \in X$, $y \in Y$.

Отображение $Q\colon D^N \to \Lambda$ называется методом (алгоритмом) построения решающих функций.

Метод бустинга

Метод бустинга предусматривает построение решения в виде

$$g(x) = \sigma \left(u_0 + \sum_{j=1}^{n} u_j s_j(x_j) + \sum_{j,k} u_{jk} s_{jk}(x_j, x_k) + \right.$$

$$\left. + \sum_{j,k,l} u_{jkl} s_{jkl}(x_j, x_k, x_l) + \ldots \right).$$

Здесь $\sigma(\cdot)$ – логистическая функция.

Модель напоминает ряд Бахадура, который даёт возможность учитывать зависимости между переменными, последовательно добавляя парные зависимости, зависимости в тройках и т.д.

Линейные методы и бустинг

Есть очевидное сходство между бустингом и логистической регрессией (бустинг на «пнях» есть попросту разновидность логистической регрессии).

Открытый вопрос: существует ли функция ядра, при которой логистическая регрессия (SVM) совпадает с методом бустинга (на деревьях).

Исследуемые методы построения решающих функций

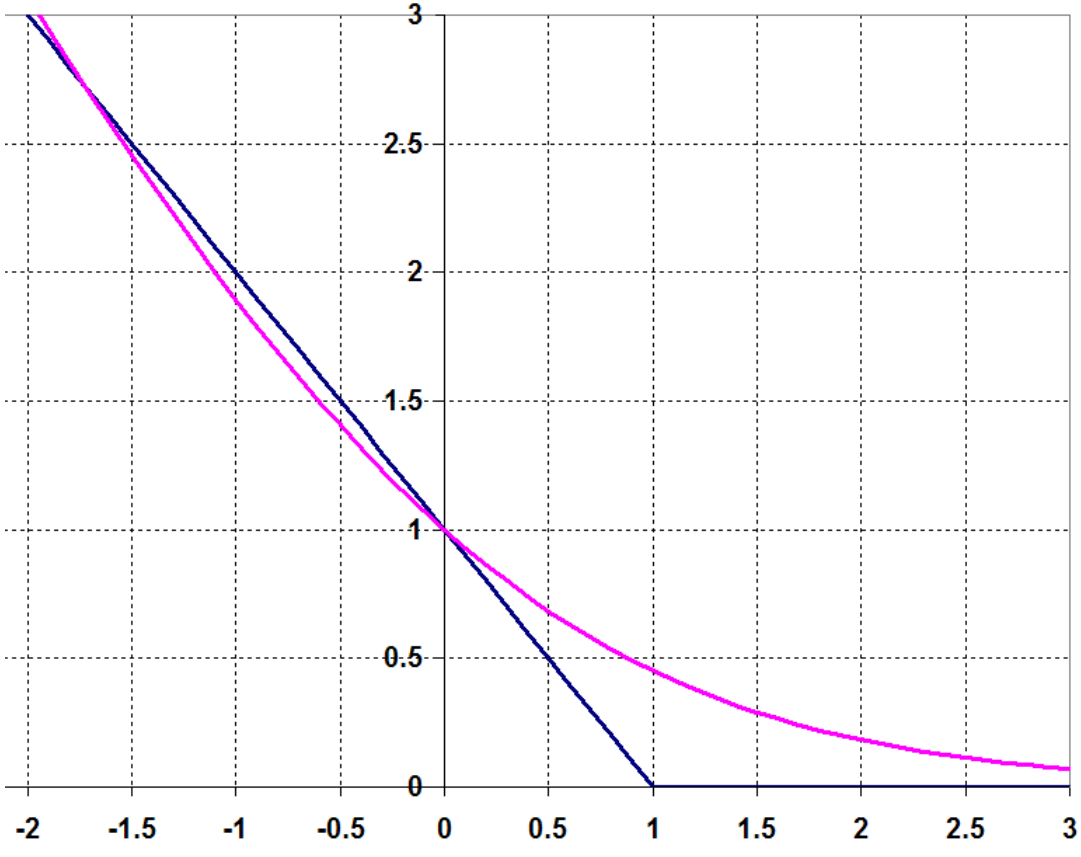
Выбраны методы:

- дискриминант Фишера;
- логистическая регрессия;
- машина опорных векторов (SVM).

Все методы линейны и (с некоторым исключением в виде логистической регрессии) основаны на геометрическом подходе.

SVM и логистическая регрессия

Отличие SVM от логистической регрессии:



Эмпирические функции потерь.

Общность методов

Методы SVM, логистической регрессии и дискриминант Фишера:

- допускают kernel trick,
- позволяют ввести регуляризатор, аналогичный величине «зазора».

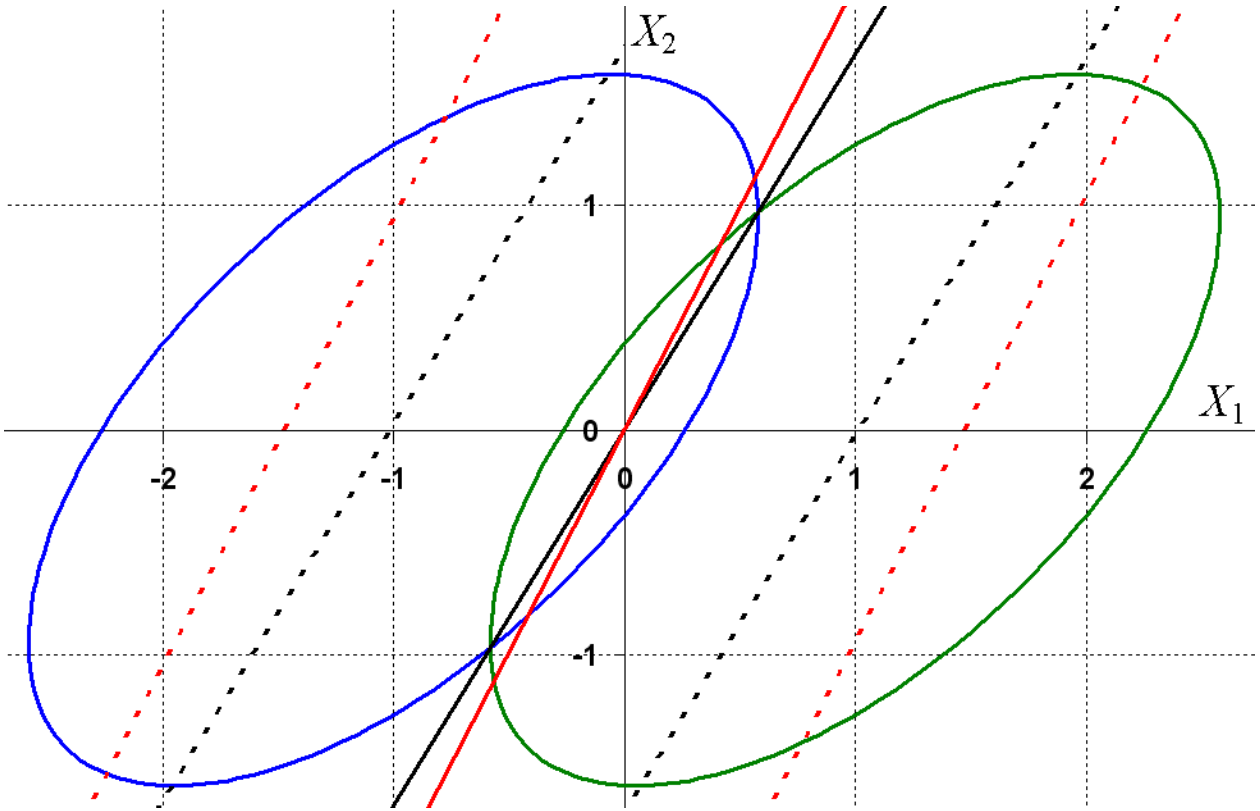
К сожалению, наиболее распространённые реализации логистической регрессии и дискриминанта Фишера не позволяют использовать ядра.

Построение решающих функций по распределениям

Применение метода построения решающих функций непосредственно к распределениям позволяет, в частности отделить погрешность аппроксимации от статистической ошибки, а также выявить некоторые свойства метода.

- При применении метода SVM к модели нормальных распределений с равными матрицами величина зазора (ширина разделяющей полосы) уменьшается при удалении распределений друг от друга.
- Существуют вероятностные модели, для которых Байесовская разделяющая функция линейна, но решения, полученные (на распределениях) методами SVM, логистической регрессии и дискриминанта Фишера не являются оптимальными.

SVM на нормальных распределениях



Параметр регуляризации огрубляет решение.

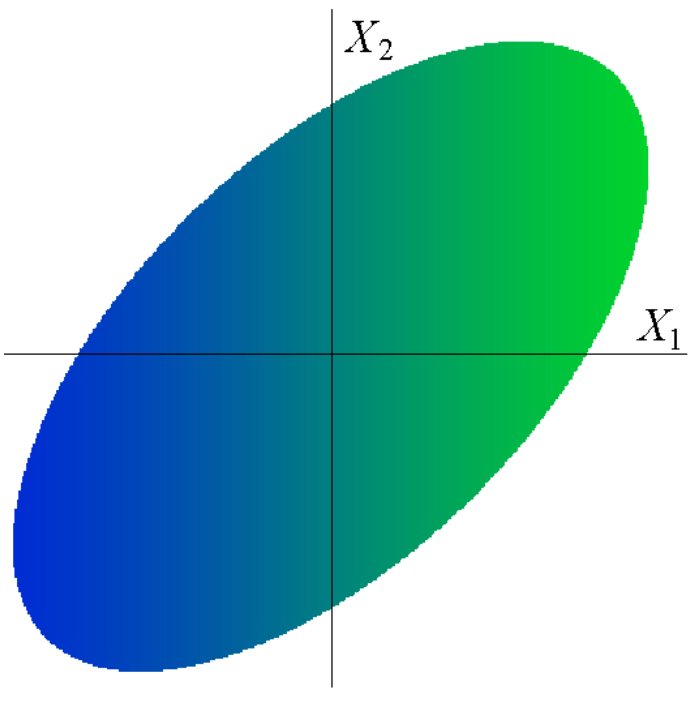
Модели максимального правдоподобия

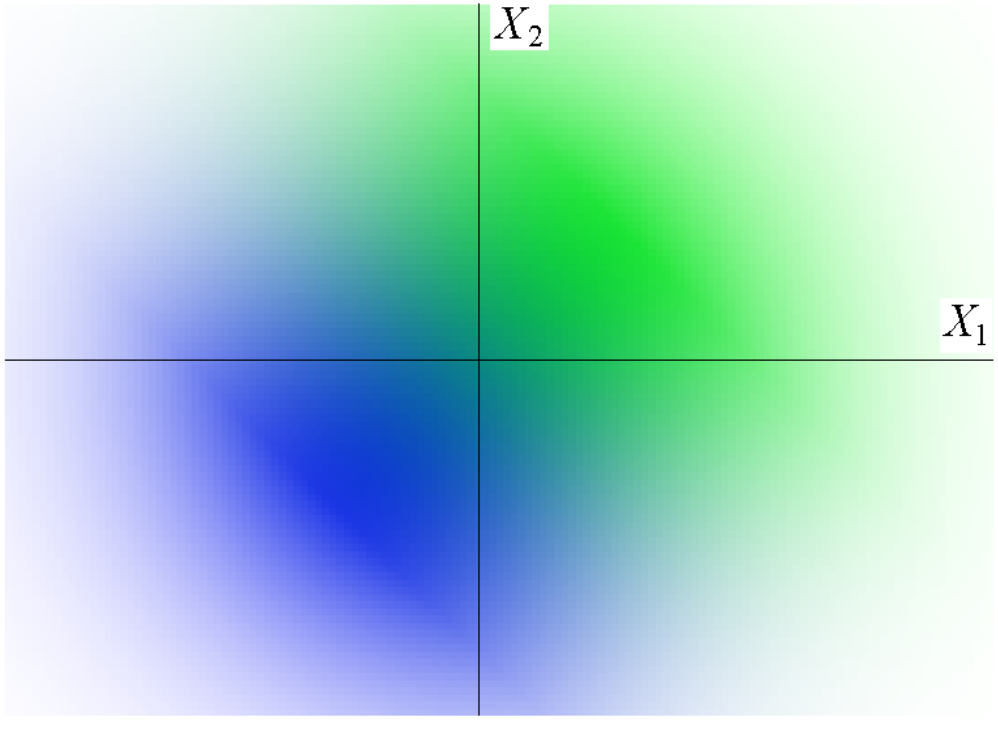
Текущая цель — построить для каждого метода вероятностную модель, на которой он был бы лучшим (решение уже строим по выборкам).

- Логистическая регрессия строится как метод максимального правдоподобия на условных распределениях.
- Для SVM модель максимального правдоподобия включает безусловное распределение.
- Модель нормальных распределений с равными ковариационными матрицами не является ММП для дискриминанта Фишера.

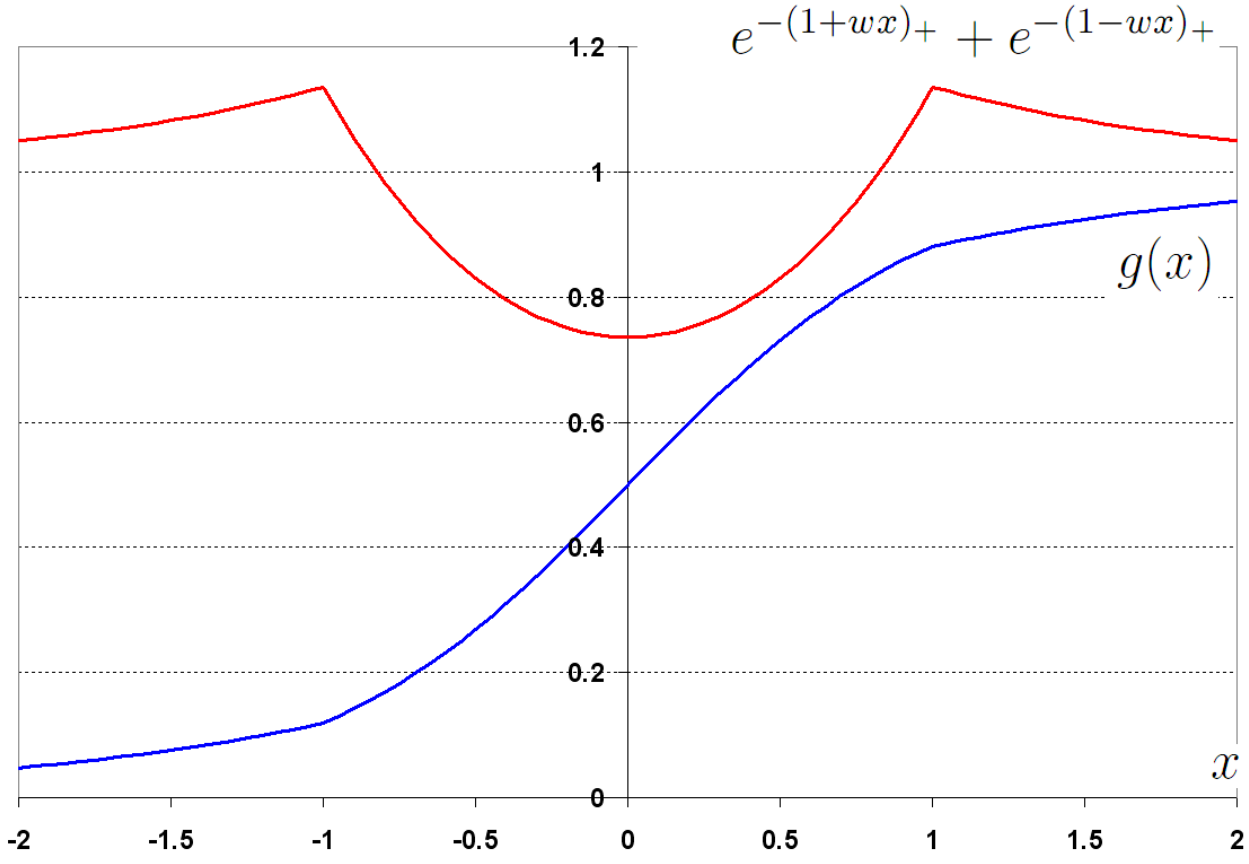
Вероятностная модель максимального правдоподобия должна иметь то же число свободных параметров, что и метод.

Модели для логистической регрессии и SVM





Модель для SVM



Компонента безусловной плотности и функция условной вероятности.

Тестовые вероятностные модели

В первую очередь, для каждого метода была сконструирована вероятностная модель, на которой этот метод предположительно должен давать наилучший результат.
Для дискриминанта Фишера это модель с нормальными распределениями, для SVM и логистической регрессии это модели максимального правдоподобия.

Предварительный численный эксперимент, однако, показал, что на всех трёх моделях лучшим оказывается дискриминант Фишера. В связи с этим были целенаправлено подобраны модели, позволяющие каждому методу продемонстрировать преимущество.

Качество решений на различных моделях

Вероятностная модель	Методы классификации				
	Байес.	ЛДФ	Логист. рег.	SVM, $\kappa = 0$	SVM, $\kappa = 0{,}2$
1. Норм. расп.	0,216	0,235	0,237	0,241	0,259
2. Норм., шум	0,244	0,313	0,309	0,305	0,326
3. Логист.	0,345	0,380	0,382	0,389	0,428
4. Лог., шум	0,359	0,433	0,430	0,431	0,464
5. Лог., смесь	0,320	0,365	0,352	0,364	0,414
6. Логист.	0,202	0,220	0,223	0,226	0,256
7. Правд. SVM	0,207	0,228	0,232	0,233	0,258

Выводы

- Для каждого из методов удалось построить модель, на которой этот метод лучший, но это не модели максимального правдоподобия.
- Дискриминант Фишера превосходит методы SVM и логистической регрессии на многих моделях с распределениями, далёкими от нормального.
- Перспективна разработка робастной версии дискриминанта Фишера с ядрами.
- Открытый вопрос: существует ли функция ядра, при которой логистическая регрессия (SVM) совпадёт с методом бустинга (на деревьях).