Об оптимальности методов построения решающих функций

Неделько В. М.

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

Спецкурс «Теория статистических решений».
Лекция 8.

Основные понятия

Пусть X – пространство значений переменных, используемых для прогноза,
$Y = \{-1, 1\}$ – 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.
Параметр c будем называть *стратегией природы*.

Риск

Решающей функцией (алгоритмом классификации) называется соответствие $\lambda\colon X \to Y$.
Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}\colon Y^2 \to [0, \infty)$.
Например, $\mathcal{L}(y, y') = \begin{cases} 0, & y{=}y' \\ 1, & y{\neq}y' \end{cases}$.

Под риском будем понимать средние потери:

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

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

Цель — построить решающую функцию, которая бы минимизировала риск, но риск зависит от распределения, которое неизвестно.

Метод построения решающих функций

Пусть $Q\colon D^N \to \Lambda$ — метод (алгоритм) построения решающих функций, $\lambda_{Q,V}$ — функция, построенная по выборке V методом Q, Λ — заданный класс решающих функций.

Требуется найти метод, который бы минимизировал средний риск:

$$F(c, Q) = \mathsf{E}_{D^N} R(c, \lambda_{Q,V}).$$

Неочевидным образом, эта задача оказывается близка проверке статистических гипотез.

Связь с задачей проверки гипотез

Самые первые методы машинного обучения как правило проверяли гипотезу (например нормальности), и в соответствии с ней строили решение.
Но уже Фишер предложил метод, не делающий предположений о распределении.

- Классическая статистическая гипотеза: распределение принадлежит заданному классу.
- Статистическая гипотеза в машинном обучении: распределение таково, что заданный метод построения решиющих функций обеспечивает требуемое качество.

Байесовский подход

Часто статистические подходы к машинному обучению называют байсовской теорией.

Однако настоящий байесовский подход подразумевает задание априорных вероятностей на гипотезах.

«Парадокс конвертов»

Игроку предлагается выбрать один из двух одинаковых на вид запечатанных конвертов с деньгами, причём известно, что сумма в одном из них в 10 раз больше, чем в другом. При этом игроку разрешается вскрыть один конверт, после чего решить, забрать его или оставшийся запечатанным.

Пусть в первом конверте оказалось x рублей.
Если считать, что во втором конверте может быть равновероятно $10x$ или $\frac{x}{10}$, то математическое ожидание выигрыша при выборе второго конверта будет $5.05x$. Но это противоречит здравому смыслу.
Парадокс является классическим примером некорректного использования байесовского подхода.
Вероятностная модель должна допускать симуляцию.

О терминологии

Часто задачу машинного обучения формулируют как задачу поиска скрытой зависимости.
Но объективно никакой зависимости нет, есть только совместное распределение.
Задача же определяется критерием качества, например, для квадратичной функции потерь оптимальным решением будет регрессия, а для абсолютных потерь — условная медиана.

Варианты:

- регрессионный анализ — не всегда строим регрессию;
- восстановление зависимостей — объективной зависимости обычно нет;
- построение решающих функций — самый точный термин.

Критерии качества

Виды критериев (метрик) качества.

- Целевой критерий: функция потерь отражает реальные потери от ошибочно принятого решения. Мы не выбираем этот критерий, он — объективная данность.
- Критерий для обучения: может использовать значение целевой функции потерь на выборке, а может включать любую другую функцию потерь. Не обязательно вообще выражается через потери (дискриминант Фишера). Обычно включает регуляризацию.

Критерии точности классификации

Матрица ошибок. Ошибки первого и второго рода.

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

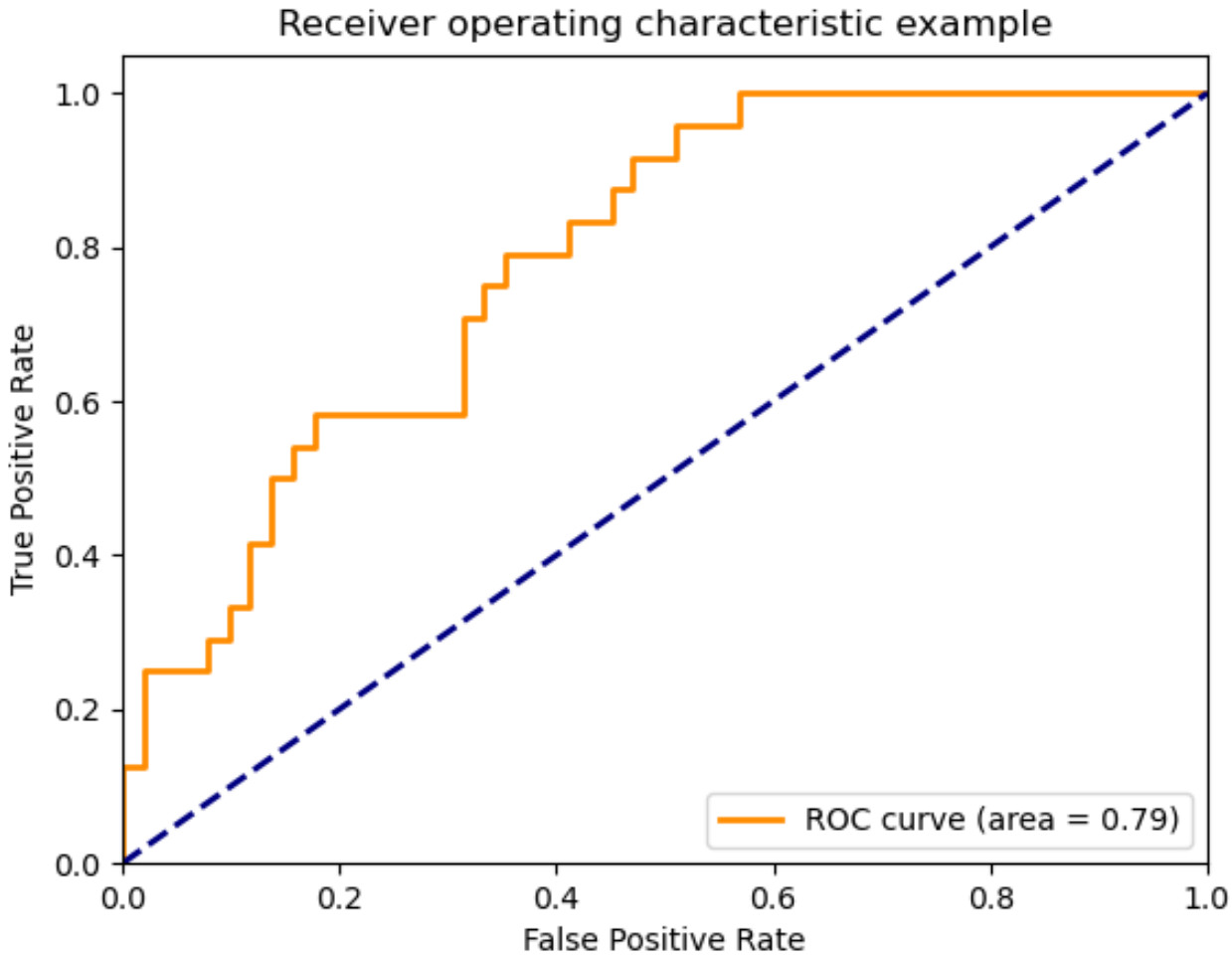
Из этих частот получаем: accuracy, precision, recall, specificity.

Также можно построить ROC-кривую и PR-кривую.

Кривая ошибок

Распространённым критерием качества оценки $\tilde{g}(x)$ является AUC — area under the curve, т.е. площадь под так называемой ROC-кривой (receiver operating characteristic, или кривая ошибок).
Пусть $F^y_{\tilde{g}}(z)$ — условная функция распределения случайной величины $\tilde{g}(x)$, определяемая условной мерой $\mathsf{P}_c(E \mid y)$, где $E \subseteq X$ — событие.
Тогда ROC-кривая определяется как кривая, заданная параметрически множеством точек $\left(F^{-1}_{\tilde{g}}(z), F^1_{\tilde{g}}(z)\right)$, когда z изменяется от $-\infty$ до $+\infty$, и отрезков, соединяющих последовательные точки в случае разрывов функций распределения.

Пример кривой ошибок



Свойства кривой ошибок

Начало ROC-кривой в точке $(0, 0)$, конец — в $(1, 1)$.

Чем больше значение AUC, тем лучше решение $\tilde{g}(x)$.

Значение 0,5 соответствует наихудшему качеству $\tilde{g}(x)$, учитывая что $\mathrm{AUC}(1 - \tilde{g}(x)) = 1 - \mathrm{AUC}(\tilde{g}(x))$.

Строго монотонное преобразование функции $\tilde{g}(x)$ не меняет AUC.

Для «обычной» решающей функции значение AUC есть просто среднее арифметическое между вероятностями правильного прогнозирования каждого класса.

На выборке используются эмпирические функции распределения.

Функция потерь для оценки вероятности

Возможна более общая постановка задачи, когда под решающей функцией понимается оценка $\tilde{g}(x)$ условной вероятности

$$g(x) = \mathsf{P}_c(y = 1 \mid x) = \frac{\mathsf{P}_c(dx, y = 1)}{\mathsf{P}_c(dx)}.$$

Качество решения $\tilde{g}(x)$ можно выражать следующей функцией потерь

$$\mathcal{L}(y, \tilde{g}(x)) = -I(y = 1) \cdot \ln \tilde{g}(x) - I(y = -1) \cdot \ln(1 - \tilde{g}(x))$$

Выборочное среднее данной функции потерь есть взятая со знаком минус функция правдоподобия выборки по отношению к оценке условной вероятности.

Критерии качества классификации

- Точность (accuracy), balanced accuracy.
- F1 score.
- ROC-AUC, Gini coefficient.
- Logloss.

Метод построения решающих функций

Пусть $Q\colon D^N \to \Lambda$ — метод (алгоритм) построения решающих функций, $\lambda_{Q,V}$ — функция, построенная по выборке V методом Q, Λ — заданный класс решающих функций.

Метод $\tilde{Q}$, минимизирующий эмпирический риск, есть

$$\lambda_{\tilde{Q},V} = \arg\min_{\lambda\in\Lambda} \widetilde{R}(V,\lambda).$$

Сравнение методов построения решающих функций

- Выбор эталонного набора тестовых задач.
- Введение понятия оптимальности метода.

Сопоставление с задачей проверки гипотез

Статистический критерий можно считать частным случаем метода построения решающих функций, когда в роли решающей функции выступает предикат.

В роли риска ошибка второго рода, но функция потерь зависит от распределения.

Известно множество статистических критериев, но нет понятия наилучшего критерия (кроме случая известной простой альтернативы).

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.

Система «Полигон» — 2000-е

Воронцов К.В., Ивахненко А.А., Инякин А.С., Лисица А.В., Минаев П.Ю. «Полигон» — распределённая система для эмпирического анализа задач и алгоритмов классификации // Всеросс. конф. Математические методы распознавания образов-14 - М.: МАКС Пресс, 2009. С. 503–506.

Принципы:

- использование реальных задач,
- большое число характеристик качества,
- основной критерий - скользящий экзамен.

Тестовые единицы

Возможные тестовые единицы:

- таблица данных,
- распределение,
- класс распределений.

Проблема определения оптимальности метода

Напомним, что метод — это отображение выборок в решения.

- Для таблицы данных понятие оптимального метода не имеет смысла.
- Для заданного распределения оптимальный метод вырожден — он любой выборке сопоставляет байесовское решающее правило.
- Об оптимальности метода можно говорить только для класса распределений.
- Даже для нормальных распределений оптимальный метод неизвестен.

Минимаксный подход к оцениванию качества

- Максимальное по классу распределений значение риска для всех методов одинаково.
- Вводить ограничения сверху на Байесовский уровень ошибки не имеет смысла.
- Использование максимума не абсолютного риска, а отнесённого к достижимому уровню, позволяет ввести осмысленное понятие метода, оптимального на классе распределений.

Похожая проблема в статистическом оценивании решается требованием несмещённости эффективной оценки.

Достижимый уровень качества

- Интересует не Байесовский уровень ошибки, а тот, который реально достижим.
- Необходимо задавать для каждого распределения из класса.
- Определяется на основе эталонного метода или из других соображений.

Выбор классов распределений

- Класс распределений не должен быть ни узким ни широким — иначе получаем соответственно вырожденный метод или аналог NFL.
- Представительность для исследований: все значения достижимого риска, максимально смещённые оценки.
- Параметр сложности. Универсальность.
- Замкнутость относительно допустимых преобразований пространства переменных.

Варианты классов распределений

- Класс кусочно–постоянных распределений.
- Класс нормальных распределений.
- Класс, сформированный случайными решающими деревьями.
- Ядерные функции для условных вероятностей.

Выводы

- Задача машинного обучения заключается в конструировании методов построения решающих функций.
- Критерий качества первичен. Он определяет оптимальное решение.
- Необходимо различать целевой и эмпирический критерии.
- До сих пор не введено понятие, что значит: один метод классификации лучше другого.
- Задачи машинного обучения похожи на задачи проверки статистических гипотез.