Выборочные функционалы качества классификации

Неделько В. М.

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

Спецкурс «Теория статистических решений».
Лекция 7.

Точечные оценки риска

- контрольная выборка,
- эмпирический риск,
- скользящий экзамен (кроссвалидация),
- out-of-bag,
- другие статистики (bootstrap, комбинации статистик).

Желательные свойства точечных оценок

- несмещённость,
- состоятельность,
- эффективность.

В отличие от оценивания параметров оценивание риска (не по контрорльной выборке) подразумевает оценивание случайной величины.

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. 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 $\mathcal{L}\colon Y^2 \to [0, \infty)$.
Положим $\mathcal{L}(y, y') = \left\{ \begin{smallmatrix} 0, & y = y' \\ 1, & y \neq y' \end{smallmatrix} \right.$.

Под риском будем понимать средние потери:

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

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

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

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

Эмпирический риск

Пусть $V = \big((x^i, y^i) \in D \,|\, i = 1, \ldots, N\big)$ – случайная независимая выборка из распределения P_c, $V \in D^N$.

Эмпирический риск определим как средние потери на выборке:

$$\widetilde{R}(V, \lambda) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^i, \lambda(x^i)).$$

Функционал, поскольку функция от (решающей) функции (и выборки).

Свойства эмпирического риска

- простота вычисления,
- сильная смещённость,
- состоятельность при соответствующем ограничении на сложность метода (оценки Вапника-Червоненкиса),
- малая дисперсия.

Контрольная выборка

Пусть $V^* = \big((x^i, y^i) \in D \,|\, i = 1, \ldots, N^*\big)$ – «новая» случайная независимая выборка из распределения P_c, $V^* \in D^{N^*}$.

Оценку риска определим как средние потери на контрольной выборке:

$$R^*(V^*, \lambda) = \frac{1}{N^*} \sum_{i=1}^{N} \mathcal{L}(y^i, \lambda(x^i)).$$

Модельный пример

Известно, что в урне белые и чёрные шары. Извлекли 10 шаров, все оказались белыми. Какой прогноз о цвете следующего шара?

Пусть p – вероятность чёрного шара

$$\mathsf{P}_p(M) = C_N^M p^M \cdot (1-p)^{N-M}, \quad \mathsf{P}_p(0) = (1-p)^N.$$

Положив $\mathsf{P}_p(0) = \alpha$, имеем $p = 1 - \alpha^{\frac{1}{N}} = 1 - e^{\frac{\ln\alpha}{N}}$.
При $\alpha = 0{,}1$ и $N = 10$ получим $p \approx 0{,}2$.

Доверительный интервал в схеме Бернулли

Односторонний интервал $[0, \hat{p}]$

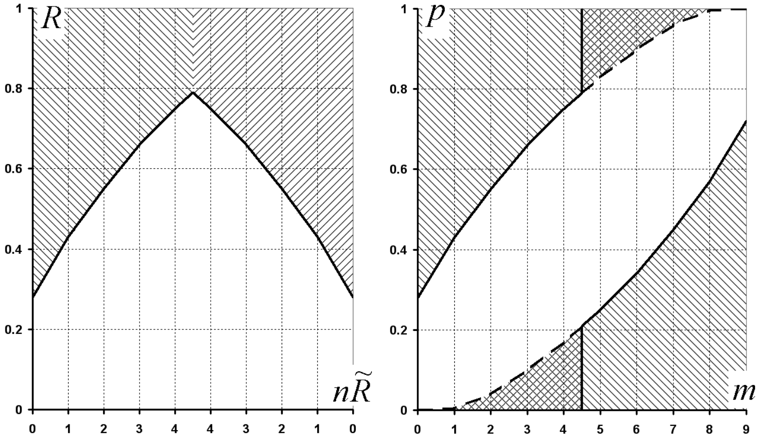
$$\sum_{i=0}^{M} C_N^i \hat{p}^i \cdot (1 - \hat{p})^{N-i} = \alpha.$$

Двусторонний интервал $[p_1, p_2]$

$$\sum_{i=0}^{M} C_N^i p_2^i \cdot (1 - p_2)^{N-i} = \sum_{i=M}^{N} C_N^i p_1^i \cdot (1 - p_1)^{N-i} = \frac{\alpha}{2}.$$

Пример критического множества

Пусть $L = 2$, $\lambda_2(x) = 1 - \lambda_1(x)$, $p = \mathsf{P}(y = 0)$, m – количество объектов $y = 0$ в выборке.



Байесовский подход

Положим равномерное $\varphi(p) \equiv 1$.
Формула Байеса

$$\varphi(p \,|\, M) = \mathsf{P}(M \,|\, p)\frac{\varphi(p)}{\mathsf{P}(M)}.$$

Используя нормировку, получаем

$$\varphi(p \,|\, M) = (N+1)C_N^M p^M \cdot (1-p)^{N-M}.$$

Можем вычислить

$$\mathsf{E}_M p = \int\limits_0^1 p\varphi(p \,|\, M)\, dp = \frac{M+1}{N+2}.$$

Усреднение по доверительной вероятности

Считаем $\eta(\hat{p}) = 1 - \alpha(\hat{p})$ функцией распределения.
Можно усреднить

$$\hat{\mathsf{E}}_M p = \int\limits_0^1 \hat{p}\, d\eta(\hat{p}) = \frac{M+1}{N+1}.$$

Если нельзя, но очень хочется, то — можно.

Свойства оценки по контрольной выборке

- простота вычисления,
- несмещённость (не совсем в том смысле, в котором хотелось бы),
- состоятельность,
- известен точный доверительный интервал,
- эффективность (не в том смысле, в котором хотелось бы),
- требуют дополнительной выборки.

Скользящий экзамен

Функционал скользящего экзамена определяется как:

$$\breve{R}(V, Q) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^i, \lambda_{Q, V_i'}(x^i)),$$

где $V_i' = V \setminus \{(x^i, y^i)\}$ — выборка, получаемая из V удалением i-го наблюдения,

Несмещённость скользящего экзамена

Теорема

$$\mathsf{E}\breve{R}(V_N, Q) = \mathsf{E}R(V_{N-1}, Q).$$

Доказательство элементарно, хотя неочевидно.

Во-первых, используется факт, что математическое ожидание суммы есть сумма мат. ожиданий в т.ч. для зависимых случайных величин.

Во-вторых, используется несмещённость оценки hold-out. А поскольку кроссвалидация — это усреднение нескольких оценок hold-out, она также получается несмещённой.

Несмещённость оценки hold-out

Оценка hold-out строится на так называемой «отложенной» выборке.
Иногда считается, что это частный случай кроссвалидации, при котором разбиение на обучающую и контрольную подвыборки делается только один раз.
Технически hold-out выглядит как оценка по контрольной выборке (и часто используется как её синоним). Однако о контрольной выборке мы говорим, если оцениваем уже построенное решение. А метод hold-out предполагает, что полученную оценку мы будем переносить на решение, котоорое будет построено по полной выборке.

Если однако использовать hold-out выглядит как оценку текущего решения, то отложенная выборка эквивалентна контрольной, поэтому оценка получается несмещённой.

Несмещённость оценки hold-out

Используются рассуждения, аналогичные рассуждениям в следующей задаче.

В урне 7 чёрных и 3 белых шара. Наугад извлекли шар. Затем извлекли второй шар, который оказался белым. Какова вероятность, что и первый шар — белый?

Решение. Поскольку при извлечении первого шара мы не посмотрели на его цвет, ответ будет таким же, как если бы этот шар извлекался вторым.

Cross-validaton

K-fold cross-validaton: исходная выборка разбивается на K равных частей (для простоты полагаем, что N кратно K).

$$\breve{R}^K(V, Q) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^i, \lambda_{Q, V_i^K}(x^i)),$$

где V_i^K — выборка, получаемая из V удалением всей подвыборки, которой принадлежит i-е наблюдение.

Разновидности скользящего экзамена

- leave–one–out,
- k-fold crossvalidation,
- случайные подвыборки,
- со стратификацией (выравнивание частот классов по фолдам).

На практике различие в точности несущественно, поэтому достаточно использовать, например, 5-фолд (без стратификации).

Дисперсия оценки практически не зависит от числа фолдов (для детерминированных методов).

Свойства оценки скользящего экзамена

- относительная простота вычисления,
- несмещённость (не в том смысле, в котором хотелось бы),
- состоятельность (не доказана?),
- большая дисперсия, нет приемлемых оценок доверительного интервала для риска (есть эмпирические свидетельства, что точность сравнима с контролем половинной длины),
- вычислять разброс по фолдам не имеет практического смысла.

Оценка bootstrap

Оценка bootstrap есть

$$\check{R}(V,Q) = \frac{1}{\mathsf{E}|J_0|}\mathsf{E}\sum_{i\in J_0}\mathcal{L}(y^i, \lambda_{Q,\acute{V}}(x^i)),$$

где $\acute{V}$ – выборка, получаемая из V путем N-кратного случайного (равновероятного) выбора ее значений с повторениями, J_0 – множество индексов объектов из V, ни разу не выбранных в $\acute{V}$, математическое ожидание подразумевает усреднение по выборкам $\acute{V}$.
Ввиду того, что оценка bootstrap является смещенной, чаще используют ее в комбинации с эмпирическим риском

$$\ddot{R}(V,Q) = e^{-1}\widetilde{R}(V,Q) + (1-e^{-1})\check{R}(V,Q).$$

Свойства оценки bootstrap

- относительная высокая трудоёмкость вычисления,
- приблизительная несмещённость,
- состоятельность (не доказана?),
- дисперсия неизвестна, но, вероятно, сопоставима со скользящим экзаменом.

Оценка out-of-bag

Используя кроссвалидацию (равно как и bootstrap), мы несколькими способами разбиваем выборку на обучения и контроль, и для каждого разбиения строим решающую функцию.
Далее предполагается, что в качестве итогового решения мы построим новую решающую функцию, уже по всей выборке.

Однако мы можем в качестве итогового решения усреднить уже построенные решающие функции.

Таким образом, оценка out-of-bag — это та же кроссвалидация (или bootstrap), но с другим финальным решением.
И свойства у неё уже не совсем как у кроссвалидации.

Свойства оценки out-of-bag

- относительная высокая трудоёмкость вычисления («бесплатна» для RandomForest),
- особенно актуальна для нейронных сетей (помимо RandomForest),
- приблизительная несмещённость (возможно, несколько пессимистична),
- состоятельность (не доказана?),
- дисперсия неизвестна, но, вероятно, сопоставима со скользящим экзаменом.

Гистограммный классификатор

Пусть $X = \{1, ..., k\}$. Тогда вероятностная мера $P_c[D]$, $c \in C$, задается набором вероятностей

$$\alpha_j = \mathsf{P}(x = j), \;\; p_j = \mathsf{P}(y = 0 \,|\, x = j).$$

Выборка представляется совокупностью пар

$$V = (v_j | j = \overline{1, k}), \;\; v_j = (m_j, n_j).$$

Решающая функция минимизирует эмпирический риск независимо в каждой точке $x \in X$: $f(x) = I(m_j < n_j)$.

Выражения для риска

$$\widetilde{R}(V) = \sum_{j=1}^{k} \tilde{r}(m_j, n_j),$$

$$\tilde{r}(m, n) = \frac{1}{N}\tilde{\nu}(m, n), \quad \tilde{\nu}(m, n) = \min(m, n - m);$$

$$R(c, \tilde{\lambda}_{Q,V}) = \sum_{j=1}^{k} r(m_j, n_j, \alpha_j, p_j),$$

$$r(m, n, \alpha, p) = \alpha\nu(m, n, p),$$

$$\nu(m, n, p) = \begin{cases} 1 - p, & m > n - m; \\ p, & m < n - m; \\ 0{,}5, & m = n - m. \end{cases}$$

Качество оценок

В общем случае оценочный функционал — это некоторая функция выборки.

Качество эмпирического функционала $\overline{R}(V,Q)$ как оценки риска обычно характеризуют средним квадратом уклонения, т.е.

$$\Delta = \mathsf{E}\left(\bar{R}(V,Q) - R(c,\lambda_{Q,V})\right)^2.$$

Существенная проблема заключается в том, что выражения зависят от c – распределения, которое неизвестно.
Кроме того, одно и то же отклонение при разных значениях риска имеет разную значимость.

Доверительный интервал для риска

Доверительный интервал для R зададим в виде $[0, \hat{R}(V)]$, где $\hat{R}(V)$ – оценочная функция или просто оценка (риска). При этом должно выполняться условие:

$$\forall c, P_c\big(R \leq \hat{R}(V)\big) \geq \eta,$$

где η – заданная доверительная вероятность.
На практике интервальную оценку будем строить как $\hat{R}(\bar{R}(V))$ –функцию точечной оценки.

Качество интервальной оценки будем характеризовать величиной $\mathsf{E}\hat{R}(V)$, которая зависит от c, в виду чего выбор наилучшей оценки становится многокритериальной задачей.

Эмпирические доверительные интервалы для риска

Эмпирический доверительный интервал для R зададим в виде $[0, \hat{R}(\bar{R}(V))]$.
При этом должно выполняться условие:

$$\forall c \in \widetilde{C}, P_c\big(R \leq \hat{R}(V)\big) \geq \eta,$$

где η – заданная доверительная вероятность, а $\widetilde{C}$ – эвристически выбранное множество распределений.

Сравнение интервалов

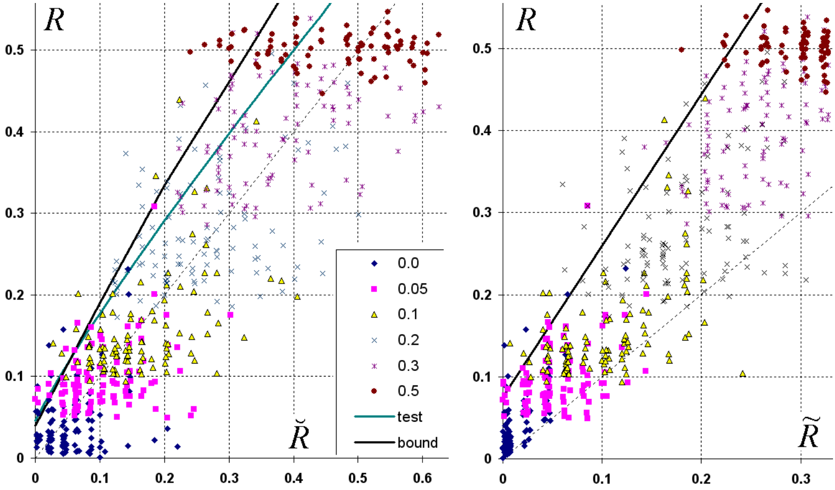
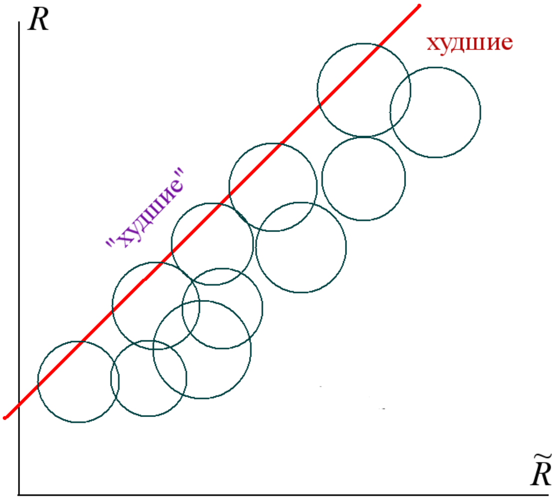


Схема оценивания риска



Замечания

- наилучший способ оценивания риска неизвестен,
- на практике обычно используют скользящий контроль,
- по обучающей выборке нет приемлемых оценок доверительного интервала для риска,
- полезно использовать статистическое моделирование.

«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.