Исследование погрешности оценок скользящего экзамена*

Неделько В. М.
nedelko@math.nsc.ru
Новосибирск, Институт математики СО РАН

В работе на модельных задачах проводится сравнение различных вариантов оценки скользящего экзамена, таких как leave-one-out и K-fold cross-validaton, а также оценки, основанной на эмпирическом риске с поправкой на смещение. Приводятся зависимости точности оценок от байесовского уровня ошибок. В качестве методов классификации рассмотрены дискриминант Фишера и гистограммный классификатор. В рамках исследования рассмотренные оценки риска показали достаточно близкие результаты по точности.

Ключевые слова: *распознавание образов, машинное обучение, решающая функция, вероятность ошибочной классификации, эмпирический риск, скользящий экзамен.*

Investigation of accuracy of crossvalidation*

Nedelko V. M.
Institute of Mathematics SB RAS

Several kinds of cross-validation, such as leave-one-out and K-fold cross-validaton, and also an empirical risk based bias adjusted estimate are investigated. The accuracies of the estimates on synthetic data are compared using Fisher's discriminant and histogram classifier. All estimates under consideration have shown similar accuracy.

Keywords: *pattern recognition, machine learning, decision function, misclassification probability, empirical risk, crossvalidation.*

Введение

Оценка скользящего экзамена является, пожалуй, наиболее часто используемой на практике оценкой вероятности ошибочной классификации [1, 2, 3, 4]. При этом существуют различные варианты этой оценки: leave-one-out, K-fold cross-validaton, случайные разбиения выборки, разбиения с соблюдением пропорции классов и другие.

Несмотря на активные исследования в этой области, до сих пор нет полного понимания, какая из оценок предпочтительнее и в каких условиях [5].

Данная работа посвящена исследованию вопроса, какую долю объектов следует выделять для контроля. Для наглядности будут рассмотрены два варианта: leave-one-out, когда для контроля выделяется один объект, и K-fold cross-validaton, когда выборка разбивается на K частей, каждая из которых поочередно выступает в качестве контрольной выборки. На самом деле, первый вариант является частным случаем второго при $K = N$, поэтому речь идет о выборе K.

В литературе можно встретить утверждение, что наилучшим выбором является использование K от 5 до 10 и что такой вариант предпочтительнее, чем leave-one-out. Действительно, есть ряд доводов в пользу такого утверждения. Первый довод состоит в том, что leave-one-out менее устойчива. Легко привести примеры выборок для случая двух

Работа выполнена при финансовой поддержке РФФИ, проект № 11-07-00346-а.

классов, где эта оценка равна 1. Однако при малых K оценка cross-validaton становится существенно смещенной (поскольку строится при меньшем объеме обучающей подвыборки). Другой довод состоит в том, что при $K << N$ мы имеем оценку доверительного интервала для риска. Это классический доверительный интервал для параметра биномиального распределения, который применим для оценивания риска по одной контрольной выборке.

Очевидно, что оценка K-fold cross-validaton не хуже, чем оценка по одной контрольной выборке объема $\frac{N}{K}$. Фактически же оценка cross-validaton существенно лучше последней, но нет приемлемых аналитических оценок, насколько она лучше. Получается, что на практике мы рассчитываем на то, что оценка cross-validaton точнее, чем это удается доказать. Для leave-one-out вообще неизвестно приемлемых оценок точности. Но то, что мы не можем оценить ее точность, не означает, что она менее точна, чем cross-validaton.

В данной работе мы будем сравнивать точность оценок риска методом статистического моделирования.

При этом возникает еще одна нетривиальная проблема: выбор функции, характеризующей погрешность. Очевидно, что оценить вероятность 0,5 с погрешностью 0,1 совсем не то же самое, что оценить с той же погрешностью вероятность, к примеру, 0,15. Однако это соображение не позволяет однозначно определить выбор этой функции. В работе будет использован достаточно очевидный вариант функции погрешности, учитывающий приведенное соображение.

Основные понятия

Пусть X — пространство значений переменных, используемых для прогноза, а $Y = \{-1, 1\}$ — пространство значений прогнозируемых переменных, и пусть 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.

Решающей функцией 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:

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

При $\mathcal{L}(y, y') = I(y = y')$, где $I(\cdot)$ – индикаторная функция (равна 1, если условие истинно, и 0 — иначе), риск есть вероятность ошибочной классификации.

Заметим, что значение риска зависит от c — распределения, которое неизвестно. Поэтому возникает задача оценивания риска по выборке.

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

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

Наиболее просто можно оценить риск по контрольной выборке:

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

где $V^* = \big((x^i_*, y^i_*) \in D \,|\, i = 1, \ldots, N^*\big)$, $V^* \in D^{N^*}$ — отличная от V выборка из распределения P_c. В этом случае доверительный интервал для риска есть классический односторонний доверительный интервал для параметра биномиального распределения. Такой интервал не зависит от метода классификации Q.

Для оценки риска естественно использовать эмпирические функционалы качества, т.е. точечные оценки риска, такие как эмпирический риск, оценка скользящего экзамена, оценка bootstrap и т.п.

Эмпирический риск определяется как средние потери на выборке:

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

Оценка 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 $V'_i = V \backslash (x^i, y^i)$ — выборка, получаемая из V удалением i-го наблюдения,

Для вычисления оценки 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.

Функция погрешности оценки

Мерой качества точечных оценок обычно выступает средний квадрат отклонения от оцениваемой величины. Однако для оценивания риска такая функция погрешности мало подходит. Действительно, ситуация, когда мы оценили риск как нулевой, а он равен 0,1, совсем не равноценна ситуации, когда мы дали оценку 0,4 при истинном значении 0,5. В первом случае погрешность значительная, а во втором — несущественная.

Очевидным решением является использование не квадрата отклонения, а некоторой более сложной функции погрешности $\Delta(\cdot, \cdot)$.

Из содержательного смысла величины данная функция должна «усиливать» погрешность при приближении оцениваемой величины к крайним значениям 0 и 1. Однако это соображение не устраняет произвол в выборе конкретного вида функции.

Попытаемся уменьшить неопределенность, вводя дополнительные ограничения и требования.

Предположим, что мы (в схеме Бернулли) оцениваем вероятность p события по его частоте ν. Достаточно естественным представляется требование, чтобы мера качества такой частотной оценки не зависела от оцениваемой вероятности.

Это достигается выбором функции погрешности

$$\Delta_1(\bar{R}, R) = \frac{(\bar{R} - R)^2}{R(1 - R)}.$$

Рис. 1: Зависимости ожидаемой погрешности частотной оценки от вероятности события в схеме Бернулли при $N = 20$

Действительно, учитывая, что $\mathsf{E}(\nu - p)^2 = \mathsf{E}(\nu - \mathsf{E}\nu)^2 = \mathsf{D}\nu$, получаем $\mathsf{E}\Delta_1(\nu, p) \equiv \frac{1}{N}$. Таким образом, функция Δ_1 просто нормирует погрешность на дисперсию оценки.

Схожее поведение имеет функция

$$\Delta_2(\bar{R}, R) = -\bar{R} \ln \frac{\bar{R}}{R} - (1 - \bar{R}) \ln \frac{1 - \bar{R}}{1 - R}.$$

Это есть дивергенция между эмпирическим распределением и фактическим распределением, что совпадает (с точностью до знака) с функцией правдоподобия.

Несмотря на то, что приведенные функции выглядят разумными способами выражения погрешности и имеют содержательное обоснование, они обладают неподходящим свойством: когда риск принимает крайние значения (0 и 1), погрешность равна бесконечности при любом значении оценки, отличающемся от истинного. В то же время, если оценка принимает крайние значения, а сама величина — не крайние, то функция погрешности конечна.

С точки зрения свойств, более подходящим вариантом функции погрешности были бы $\Delta_1'(\bar{R}, R) = \Delta_1(R, \bar{R})$ и $\Delta_2'(\bar{R}, R) = \Delta_2(R, \bar{R})$, т.е. функции с переставленными местами аргументами. Последние варианты, правда, фактически «запрещают» оценкам принимать крайние значения, но это приемлемо. Например, если скорректировать частотную оценку, используя $\frac{N\nu+1}{N+2}$, то средняя погрешность имеет приемлемый вид.

В данной работе для сравнения оценок будет использована функция погрешности

$$\Delta_0(\bar{R}, R) = \frac{(\bar{R} - R)^2}{2\bar{R}(1 - \bar{R}) + 2R(1 - R)}.$$

Этот вариант функции погрешности имеет подходящее качественное поведение, что иллюстрирует рис. 4. На правом графике функция гораздо ближе к константной, чем на левом.

Для сравнения вариантов оценки скользящего экзамена будет решаться модельная задача методом дискриминанта Фишера.

Дискриминант Фишера

Дискриминант Фишера использует исключительно метрические свойства конфигурации выборочных точек и не требует не только никаких предположений о распределениях, но и вообще статистической постановки задачи классификации.

Идея дискриминанта Фишера заключается в выборе такого направления в пространстве переменных, при проецировании выборки на которое образы классов оказываются в некотором смысле наиболее удаленными друг от друга. Формально это выражается в максимизации следующего критерия

$$\Phi(w) = \frac{(\tilde{\mu}_{w,1} - \tilde{\mu}_{w,-1})^2}{\tilde{S}_{w,1} + \tilde{S}_{w,-1}},$$

где $\tilde{\mu}_{w,y} = \frac{1}{N_y}\sum_{i=1}^{N} wx^i \cdot I(y^i = y)$ — среднее, а $\tilde{S}_{w,y} = \frac{1}{N_y}\sum_{i=1}^{N}(wx^i - \tilde{\mu}_{w,y})^2 \cdot I(y^i = y)$ — средний квадрат отклонений проекций точек выборки y-го класса на направление w, N_y — число объектов y-го класса в выборке.

Критерий приводится к виду

$$\Phi(w) = \frac{(w\tilde{\mu}_1 - w\tilde{\mu}_{-1})^2}{w^T(\tilde{S}_1 + \tilde{S}_{-1})w} = \frac{w^T(\tilde{\mu}_1 - \tilde{\mu}_{-1})(\tilde{\mu}_1 - \tilde{\mu}_{-1})^T w}{w^T \tilde{S} w},$$

где $\tilde{\mu}_y = \frac{1}{N_y}\sum_{i=1}^{N} x^i \cdot I(y^i = y)$ — среднее точек выборки у-го класса, $\tilde{S}_y$ — выборочная ковариационная матрица y-го класса, $\tilde{S} = \tilde{S}_1 + \tilde{S}_{-1}$.

Последняя форма критерия имеет вид отношения Релея. Известно, что максимум $\Phi(w)$ достигается при $w_\Phi = \tilde{S}^{-1}(\tilde{\mu}_1 - \tilde{\mu}_{-1})$.

Заметим, что выражение для w_Φ очень похоже на выражение для нормали к разделяющей гиперплоскости для случая нормальных распределений с равными матрицами ковариаций. Отличие лишь в том, что во втором случае вместо $\tilde{S}$ используется $\frac{1}{N}(N_1\tilde{S}_1 + + N_{-1}\tilde{S}_{-1})$.

Если количество объектов обоих классов в выборке одинаково, то направления в обоих методах совпадают. Такое сходство приводит к тому, что в литературе эти методы иногда смешиваются, несмотря на их принципиальное различие по подходу и предположениям.

После выбора направления w_Φ задача классификации становится одномерной и может быть достаточно легко решена, например, выбором границы по критерию минимизации эмпирического риска. Более того, в некоторых случаях полученное (проецированием) упорядочивание объектов само по себе может считаться решением, в частности, по нему можно вычислить AUC.

Численный эксперимент

Для сравнения точности различных вариантов скользящего экзамена проведем моделирование на следующем классе распределений. Пусть X — двумерное евклидово пространство. Безусловное распределение в X подчиняется нормальному закону с нулевым средним и единичной ковариационной матрицей. Условная вероятность $\mathsf{P}(y = 1 \,|\, x)$ имеет вид логистической функции $\frac{1}{1+e^{-\alpha w_1 x}}$, где $w_1 = (1, 1)$, а α — параметр, связанный с байесовским уровнем ошибки R_0.

Для иллюстрации на рис. 2 приведена выборка из распределения данного семейства.

Дискриминат Фишера обладает высокой статистической устойчивостью, поскольку требует оценивания по выборке всего лишь порядка n параметров (n — размерность пространства), поэтому моделирование следует проводить при относительно малых объемах выборки.

Рис. 2: Выборка из моделируемого распределения

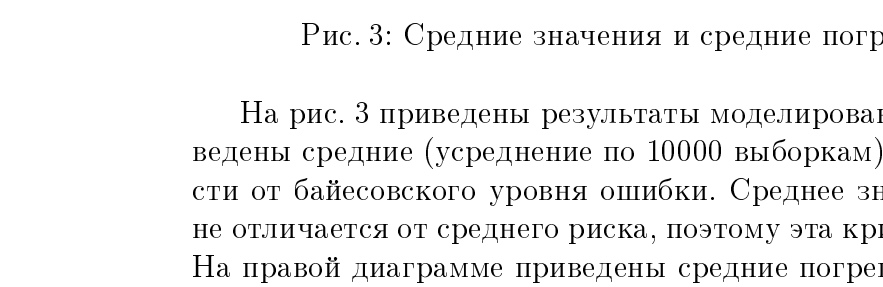

Рис. 3: Средние значения и средние погрешности различных оценок риска

На рис. 3 приведены результаты моделирования при $N = 20$. На левой диаграмме приведены средние (усреднение по 10000 выборкам) значения риска и его оценок в зависимости от байесовского уровня ошибки. Среднее значение оценки leave-one-out практически не отличается от среднего риска, поэтому эта кривая на данной диаграмме не изображена. На правой диаграмме приведены средние погрешности оценок риска.

В проводимое сравнение включена также $\hat{R}$ — оценка риска, получаемая добавлением к эмпирическому риску оценки максимального смещения [6]. Данная оценка подробно описана в [7].

Видим, что оценки leave-one-out и 5-fold cross-validation получились практически равноценными по качеству. Оценка $\hat{R}$ выглядит более предпочтительной при больших значениях R_0. Результаты моделирования при $N = 30$ и $N = 50$ качественно не отличались от приведенных.

Аналогичное исследование проводилось для гистограммного классификатора. В этом случае вычисления производились точно (не на основе статистического моделирования),

Рис. 4: Средние погрешности некоторых оценок риска для гистограммного классификатора при различных распределениях

однако ввиду экспоненциальной сложности численного суммирования результаты получены при малых размерностях: число «ячеек» равно 3, $N = 12$, $K = 4$.

На рис. 4 каждая точка соответствует некоторому распределению в дискретном пространстве. Распределения перебирались «по сетке» с достаточно большим шагом (ввиду экспоненциальной сложности). Полученный результат качественно согласуется с полученным для дискриминанта Фишера.

Заключение

В отсутствие теоретических результатов, которые бы давали точные оценки качества, важным инструментом исследования и сравнения оценок является статистическое моделирование. Хотя получаемые таким образом результаты носят частный характер, они дают возможность эмпирически делать определенные выводы.

Исследование, проведенное в данной работе, свидетельствует о практической равноценности оценок leave-one-out и 5-fold cross-validation. Хотя первая оценка в большинстве испытаний оказывается слегка точнее, в ряде случаев она дает существенно большую погрешность. На практике этот эффект, однако, вряд ли является существенным.

Также можно сделать вывод о перспективности оценки на основе эмпирического риска с поправкой на смещение. Данная оценка имеет сравнимую точность, но гораздо меньшую трудоемкость в вычислительном плане.

Литература

[1] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Институт математики СО РАН, 1999. 211 с.

[2] *Langford J.* Quantitatively tight sample complexity bounds. Carnegie Mellon Thesis. 2002. `http://citeseer.ist.psu.edu/langford02quantitatively.html`. 130 p.

[3] *Nedelko V. M.* Estimating a Quality of Decision Function by Empirical Risk // LNAI 2734. 3rd Conference (International) on Machine Learning and Data Mining in Pattern Recognition (MLDM 2003) Proceedings. Leipzig: Springer-Verlag. 2003. Pp. 182–187.

[4] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. 2008. Vol. 18, no. 2. Pp. 243–259.

[5] *Efron B., Tibshirani R.* Improvements on cross-validation: The .632+ Bootstrap method // J. Amer. Stat. Association. 1997. Vol. 92, no. 438. Pp. 548–560.

[6] *Неделько В. М.* Об интервальном оценивании риска для решающей функции // Таврический вестник информатики и математики. Изд-во НАН Украины. 2008. № 2. С. 97–103.

[7] *Недельko В. М.* Точные и эмпирические оценки вероятности ошибочной классификации // Научный вестник НГТУ. 2011. № 1 (42). С. 3–16.