Майнор «Интеллектуальный анализ данных». Семинар 15. Решение задач по деревьям, композициям и метрическим методам.

Надежда Чиркова

6 июня 2016 г.

1 Обзор изученного

В курсе были подробно изучены две задачи обучения с учителем: задача классификации и задача регрессии, а также основные подходы к их решению. Систематизируем матриал:

Группа методов	**Метод**	**Бинарная кл-я**	**Многоклассовой кл-я**	**Регрессия**
Линейные	Линейная регрессия	-	-	+
	Лог. регрессия	+	+ (one-vs-one, ove-vs-all)	-
Байесовские	Байесовский классификатор	+	+	-
Метрические	kNN	+	+	+
Деревья	Деревья	+	+	+
Композиции	Случайные леса	+	+	+
	Бустинг	+	+	+
	Беггинг	+	+	+

Решим несколько задач по последним трем темам.

2 Задачи

2.1 Деревья

Как рассказывалось на лекции, решающее дерево в каждой своей вершине m делит выборку X^m, которая оказалась в этой вершине, на две: X^l и X^r, согласно некоторому правилу. Этим правилом обычно вступает пороговая функция:

$$x^j > t?$$

x^j — j-й признак, t — порог. Чтобы разделить выборку, нужно выбрать j и t. Для этого надо перебрать все варианты (все признаки и все пороги для каждого признака) и вычислить для них значение критерия:

$$G(X^m, j, t) = H(X^m) - \frac{|X^l|}{|X^m|}H(X^l) - \frac{|X^r|}{|X^m|}H(X^l). \tag{1}$$

Здесь H — это некоторая функция, оценивающая, насколько неравномерно распределены классы в выборке. Ясно, что идеально делить выборку так, чтобы в каждой из двух подвыборок оказались объекты только одного класса. Поэтому функция H должна минимизироваться при таком идеальном разбиении. Обозначим p_k — доля объектов класса k в X^m. В sklearn реализованв две функции H:

- Критерий Джини: $H(X^m) = \sum_k p_k(1 - p_k)$;
- Энтропийный критерий: $H(X^m) = -\sum_k p_k \ln p_k$.

Попробуем разобраться, почему такие критерии работают и в чем их особенности.

Задача 1. Покажите, что критерий Джини можно записать в виде

$$H(X^m) = \sum_{k \neq k'} p_k p_{k'}. \tag{2}$$

Решение. Очевидно, $\sum_k p_k = 1$, так как p_k — это доли классов. Будем идти от (2) к формуле критерия Джини. Добавим в сумму недостающие слагаемые:

$$H(X^m) = \sum_{k \neq k'} p_k p_{k'} = \sum_{k,k'} p_l p_{k'} - \sum_k p_k = \big(\sum_k p_k\big)\big(\sum_{k'} p_{k'}\big) - \sum_k p_k^2 = \sum_k p_k - \sum_k p_k^2 = \sum_k p_k(1 - p_k).$$

В таком представлении лучше видно, что если распределение классов равномерное, что все величины p_k будут достаточно большими и $H(X^m)$ будет большим.

Задача 2. Покажите, что максимум энтропии достигается на равномерном распределении. *Решение.* Рассмотрим функцию $L(p) = p \ln p$. Продифференцируем:

$$L'(p) = 1 + \ln p; \quad L''(p) = \frac{1}{p} \geqslant 0.$$

Это означает, что $L(p)$ — выпуклая функция. По определению выпуклой функции для любых $y_1, \dots, y_n$ и любых $\alpha_k \geqslant 0, k = 1, \dots, n$, $\sum_k \alpha_k = 1$ выполняется

$$\sum_k \alpha_k y_k \leqslant \sum_k \alpha_k L(y_k).$$

Возьмем $y_k = p_k$ (любое распределение), $\alpha_k = \frac{1}{n}$ (n — число классов). Тогда $\sum_k \alpha_k y_k = \sum_k \frac{1}{n} p_k = \frac{1}{n}$ и

$$\frac{1}{n}\ln\frac{1}{n} \leqslant \frac{1}{n}\sum_k p_k \ln p_k,$$

или

$$-\ln\frac{1}{n} \geqslant -\sum_k p_k \ln p_k,$$

Легко увидеть, что левая часть равна энтропии равномерного распределения: $H(X^m) = -\sum_k \frac{1}{n}\ln\frac{1}{n} = -\ln\frac{1}{n}$, а правая часть — энтропии любого фиксированного распределения p_k.

Таким образом, максимум энтропии достигается на равномерном распределении, как и у критерия Джини. Вычислите самостоятельно эти два критерия на распределениях $(1, \ldots, 0)$.

Задача 3. Пусть дана задача предсказания того, что студент играет в крикет. Пусть в вершину m попала выборка из 30 студентов, из них 15 играют в крикет. Ее можно разделить тремя вариантами:

- по признаку пола, тогда в левой вершине окажутся 12 студентов (девушки), из них 4 играют в крикет, в правой — оставшиеся 18 студентов (юноши), из них 11 играют в крикет;
- по признаку класса, тогда в левой вершине окажутся десятиклассники (14 человек), среди которых 8 играют в крикет, в правой — 16 человек, 7 игроков;
- наконец, магическим образом нам известен признак, был ли студент вчера вечером дома: тех, кто был (левая вершина), 15 человек и 0 игроков, тех, кто не был — тоже 15, и все играют.

Вычислите значение $G(X^m, j, t)$ на трех разбиениях и ранжируйте разбиения по убыванию этого критерия.

Указание. Чтобы рассчитать величину критерия для одного разбиения, нужно в каждой из трех вершин (которую делят, левая и правая) вычислить p_k (в данном случае их две), затем вычислить H в каждой из трех вершин и подставить все в формулу (1). Решите задачу самостоятельно.

2.2 Композиции алгоритмов

Композиции алгоритмов — это способ создать новый классификатор или регрессор на основе нескольких базовых. В курсе рассмотрены три композиции:

- бэггинг: каждое дерево обучается на некоторой новой выборке объектов. Она получается либо вытаскиванием объектов с возвращением, либо выбором подмножества объектов.

- случайные леса: в них параллельно (а значит независимо) строится много решающих деревьев, каждое обучается на подмножестве объектов, подмножестве признаков и, более того, процесс построения дерева рандомизирован (j и t выбираются из случайно полученного подмножества);

- бустинг: здесь деревья строятся последовательно, каждое дерево исправляет ошибки предыдущих.

Обратим особое внимание, что композиции можно применять к любым алгоритмам, не обязательно к деревьям (особенно бэггинг и бустинг). Но деревья лучше всего для этого подходят, потому что их можно делать сильно разными, что хорошо при построении композиции.

Один из способов бэггинга (bagging) - бутстреп (bootstrap). Он подразумевает выбор N случайных объектов выборки с возвращением, то есть объекты могут повторяться.

Задача 4. Пусть $N = n$ — длина исходной выборки. Найдите вероятность того, что конкретный объект попадет в бутстрапированную выборку.

Решение. Вероятность того, что объект попадет в выборку при одном вытаскивании — $\frac{1}{n}$, n — число объектов выборки. Вероятность того, что не попадет — $1 - \frac{1}{n}$; не попадет ни при одном вытаскивании - $\left(1 - \frac{1}{n}\right)^n$. Наконец, искомая вероятность

$$1 - \left(1 - \frac{1}{n}\right)^n \to_{n\to\infty} 1 - \frac{1}{e}.$$

Этот результат можно трактовать как среднее число несовпадающих объектов в выборке.

2.3 Метрические алгоритмы

Метрические алгоритмы позволяют решать задачи классификации и регрессии. В наиболее общем случае для предсказания ответа на новом объекте каждому объекту выборки приписывается вес, зависящий от его номера в списке соседей нового объекта и от расстояния до нового объекта. В задаче регрессии для получения предсказания усредняются ответы на объектах обучающей выборки с найденными весами.

Ключевой момент при построении метрического алгоритма - выбор весов и задание функции расстояния между объектами. Расстояние задается по-разному для разных типов признаков, примеры были приведены на лекции. Остановимся на set-valued признаках, значения которых есть подмножества некоторого большого множества U. Наиболее популярное расстояние между множествами A и B — расстояние Жаккарда:

$$J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}.$$

Задача 5. Пусть множества A и B закодированы бинарными векторами фиксированной длины a и b. Каждому элементу вектора соответствует элемент большого множества U, и 1 ставится в том и только в том случае, если объект принадлежит множеству. Запишите выражение для коэффициента Жаккарда в таких обозначениях.

Решение. Мощность пересечения множеств:

$$|A \cap B| = \sum_i a_i b_i =< a, b >$$

Мощность объединения получается из формулы $|A \cap B| + |A \cup B| = |A| + |B|$:

$$|A \cup B| = \sum_i a_i + \sum_i b_i - \sum_i a_i b_i = \sum_i a_i^2 + \sum_i b_i^2 - \sum_i a_i b_i = ||a||^2 + ||b||^2 - < a, b > .$$

Наконец,

$$J(a, b) = 1 - \frac{< a, b >}{||a||^2 + ||b||^2 - < a, b >} = \frac{||a||^2 + ||b||^2 - 2 < a, b >}{||a||^2 + ||b||^2 - < a, b >} = \frac{< a - b, a - b >}{||a||^2 + ||b||^2 - < a, b >}.$$

Последний переход показывается просто, убедитесь в этом самостоятельно.

Это равносильный способ задания расстояния Жаккарда — на бинарном векторе. Отметим, что без вычитания из единицы выражение $\frac{|A \cap B|}{|A \cup B|}$ задает меру близости объектов, а не расстояние (в отличие от него, она тем больше, чем более похожи объекты). В Википедии приводятся другие похожие на близость Жаккарда коэффициенты.