ПРИМЕНЕНИЕ КЛАССИФИКАТОРОВ МАШИННОГО ОБУЧЕНИЯ ПРИ ОПРЕДЕЛЕНИИ КАЧЕСТВА СИГНАЛА В ПАРАМЕТРИЧЕСКИХ ТЕХНОЛОГИЧЕСКИХ ПРОЦЕССАХ

Шоймардонов Ж.З.,
магистрант БИТИ, г. Бухара, Республика Узбекистан
Научный руководитель – Ибрагимов У.М., к.п.н., доцент

В искусственном интеллекте и машинном обучении – задача разделения множества наблюдений (объектов) на группы, называемые классами, на основе анализа их формального описания. При классификации каждая единица наблюдения относится определенной группе или номинальной категории на основе некоторого качественного свойства.

Классификация является наиболее распространенной задачей в машинном обучении. Классификатор – это отображение $\hat{c} : X \to C$, где $C = \{C_1, C_2, \ldots, C_k\}$ – конечное и обычно небольшое множество меток классов. Иногда мы также будем использовать C_i для обозначения набора примеров этого класса. Мы используем специальнуж метку («шляпу»), чтобы указать, что $\hat{c}(x)$ является оценкой истинной, но неизвестной функции $c(x)$. Примеры для классификатора имеют вид $(x, c(x))$, где $x \in X$ – экземпляр, а $c(x)$ – истинный класс экземпляра. Изучение классификатора включает в себя построение функции $\hat{c}$ так, чтобы она максимально точно соответствовала c (и не только на обучающей выборке, но в идеале на весь экземпляр пространства X).

Материал и методы. В простейшем случае у нас есть только два класса, которые обычно называют положительными и отрицательными, $\oplus$ и, или +1 и -1. Двухклассовую классификацию часто называют бинарной классификацией (или концептуальным обучением, если положительный класс можно осмысленно назвать концептом). Фильтрация значений сигнала из датчика - хороший пример бинарной классификации, в которой значений сигнала условно принимается за положительный класс, а не за отрицательный (понятно, что положительный здесь не означает «хороший», потому что выходить из конечных границ интервала определения значений датчика). Другие примеры бинарной классификации включают медицинский диагноз (положительный класс здесь – наличие определенного заболевания) и обнаружение мошенничества с кредитными картами [1].

Дерево признаков на рис.1 (слева) можно превратить в классификатор, пометив каждый сигнал классом. Самый простой способ сделать это — назначить мажоритарный класс в каждом сигнале, что приведет к дереву решений на рис.1 (справа). Классификатор работает следующим образом: если сигнал содержит значение «вне интервал», оно классифицируется как не правильный сигнал (крайний правый лист); в противном случае появление допольнительного сигнала «о изменение напряжений» решает, будет ли оно помечено как не правильный сигнал или нет [2].

Из чисел на рис.1 мы можем понять, насколько хорошо работает этот классификатор. Крайний левый сигнал правильно предсказывает 40 сигналов, но также неправильно помечает 20 сигналов, которые не содержат ни значений «вне интервала», ни дополнительных сигналов «о изменение напряжений». Средний сигнал правильно классифицирует 10 сигналов со значением, но также ошибочно помечает 5 сигналов как не правильный. Тест сигнал значений «вне интервала» правильно отбирает 20 сигналных значений, а также 5 как привильнимы. В совокупности это означает, что 30 из 50 сигналов -значений классифицируются как не правильно, а 40 из 50 также являются правильнимы сигналами значениями [3;4].

Результаты и их обсуждение. Эффективность таких классификаторов можно обобщить с помощью таблицы, известной как таблица непредвиденных обстоятельств или матрица путаницы (рис.1 (слева)). В этой таблице каждая строка относится к

фактическим классам, записанным в тестовом наборе, а каждый столбец — к классам, предсказанным классификатором. Так, например, в первой строке указано, что тестовый набор содержит 50 положительных результатов, 30 из которых были предсказаны правильно, а 20 — неверно. Последний столбец и последняя строка дают суммы (т. е. суммы столбцов и строк). Суммы важны, потому что они позволяют нам оценить статистическую значимость. Например, таблица непредвиденных обстоятельств в рис.1 (справа) имеет те же суммы, но классификатор явно делает случайный выбор в отношении того, какие прогнозы являются положительными, а какие отрицательными - в результате распределение фактических положительных и отрицательных результатов в любом предсказанном классе совпадает с общим распределением (в данном случае равномерным).

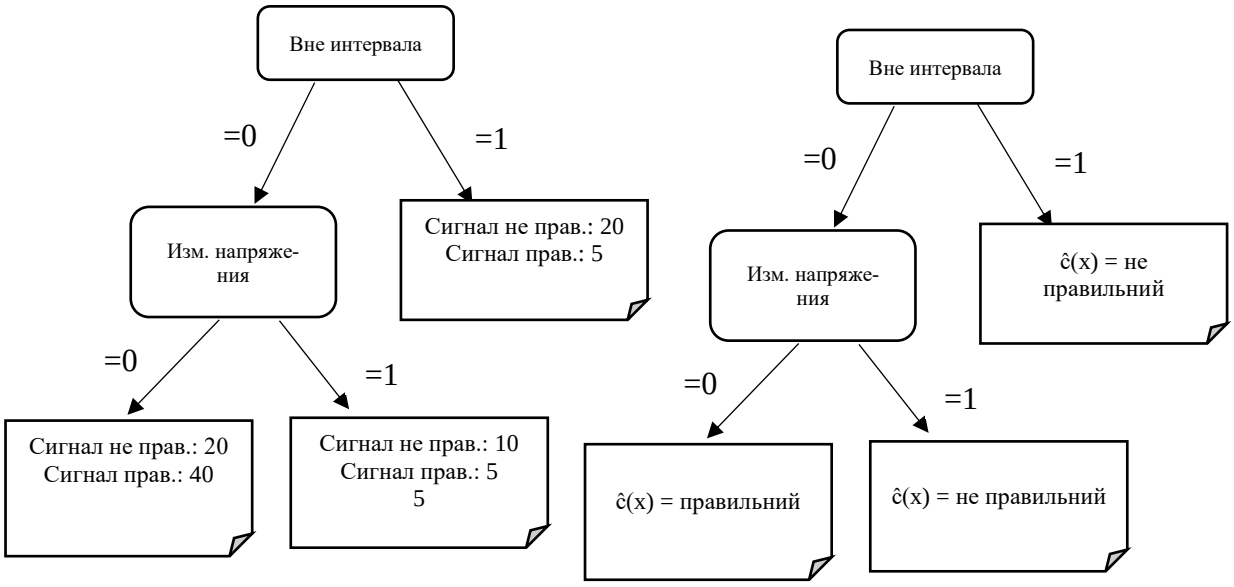

Рисунок 1. (слева) Дерево признаков с распределением классов обучающего набора в сигналах. (справа) Дерево решений, полученное с использованием решающего правила мажоритарного класса

Заключение. Точность по тестовому набору Te определяется как доступ только к истинным классам небольшой части пространства экземпляров, поэтому оценка — это все, на что мы можем надеяться. Поэтому важно, чтобы набор тестов был максимально репрезентативным. Это обычно формализуется предположением, что появление экземпляров в мире, т. е. насколько вероятно или типично конкретное сигнальное значение, управляется неизвестным распределением вероятностей на X, и что тестовый набор Te генерируется в соответствии с этим распределением.

1. Allwein, E.L., Schapire, R.E. and Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. In P. Langley (ed.), Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), pp. 9–16.

2. Blockeel, H., De Raedt, L. and Ramon, J. (1998). Top-down induction of clustering trees. In J.W. Shavlik (ed.), Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), pp. 55–63.

3. Boser, B.E., Guyon, I. and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the International Conference on Computational Learning Theory (COLT 1992), pp. 144–152.

4. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). Classification and Regression Trees. Wadsworth.