Соковнин Игорь Леонидович, факультет Искусственного интелекта

Практическое задание.

Урок 4. Оценка и интерпретация полученной модели. Обсуждение курсового проекта.

1. **Расскажите, как работает регуляризация в решающих деревьях, какие параметры мы штрафуем в данных алгоритмах?**

Регуляризацией в машинном обучении в широком смысле называется техника избегания переобучения моделей, основывающаяся на том наблюдении, что при переобучении в частности линейных моделей веса признаков становятся неоправданно большими. Идея техники заключается в том, чтобы включить в функцию потерь модели информацию об абсолютном значении весовых коэффициентов.

Некоторые виды регуляризации:

L_1-регуляризация (англ. lasso regression), или регуляризация через манхэттенское расстояние:

$$L_1 = \sum_i (y_i - y(t_i))^2 + \lambda \sum_i |a_i|.$$

L_2- регуляризация, или регуляризация Тихонова (в англоязычной литературе — ridge regression или Tikhonov regularization), для интегральных уравнений позволяет балансировать между соответствием данным и маленькой нормой решения:

$$L_2 = \sum_i (y_i - y(t_i))^2 + \lambda \sum_i {a_i}^2.$$

В качестве слагаемого регуляризации используется сумма нескольких норм с разными коэффициентами.

Наиболее часто применяемые способы регуляризации:

• Способ **Ridge** заключается в использовании нормы L2. При наличии признаков, имеющих высокий уровень корреляции между собой, данный способ стремится к выравниванию весов двух этих признаков. Данный способ штрафует большие по модулю веса строго, так как в соответствующей норме вес берется во второй степени. Напротив, малые по модулю коэффициенты штрафуются нестрого.

• Способ **Lasso** заключается в использовании нормы L1. Считается, что при таком способе признаки выбираются разреженным образом, то есть число ненулевых весов будет меньше, чем у способа Ridge. При наличии коррелированных признаков, в общем случае только один из них будет иметь ненулевой вес в итоговой модели. Этот способ штрафует большие по модулю и малые по модулю веса более однородно, чем Ridge.

• Способ **Elastic net** является компромиссом между предыдущими двумя и представляет собой использование линейной комбинации L1 и L2 норм с некоторыми независимыми коэффициентами. Данный способ стремится одновременно к

уравниванию весов и выбору разреженным образом коррелированных признаков одновременно.

В качестве регуляризации модели случайного леса обычно производится ограничение глубины строящихся решающих деревьев; при прочих равных предпочтение разбиения по признакам (штраф за размер, сложность дерева), по которым разбиение уже производилось в других деревьях леса; обрезание деревьев. Также для увеличения независимости отдельных классификаторов, а значит, для меньшего смещения в среднем, для обучения отдельных деревьев часто берется подвыборка обучающей выборки (типичные объемы — 70-90%), а также при принятии решения о разбиении множества в узле дерева рассматривается подмножество признаков (типичные размеры подмножества признаков снова 70-90%, однако могут встречаться и другие).

При обучении модели градиентного бустинга параметром регуляризации выступает коэффициент обучения, который обычно принимает значение от 0 до 1 и обозначает снижение вклада каждой новой базовой предсказывающей модели, добавляемой в общую модель градиентного бустинга на очередной итерации. При обучении модели градиентного бустинга для избежания переобучения обычно либо снижают число итераций обучения, либо уменьшают коэффициент обучения. Эти параметры имеет смысл настраивать одновременно, при этом чем больше значение одного параметра, тем меньше должно быть значение другого параметра для получения хороших результатов предсказаний модели. Как и в случае случайного леса, используют технику выбора подмножеств объектов для обучения очередного дерева и подмножества признаков для разбиения множества на два в узле дерева, если используется решающее дерево в качестве базового предсказателя.

В моделях метода опорных векторов параметром регуляризации служит параметр C, отвечающий за компромисс между шириной отступа (коридора), которую нужно максимизировать и количеством объектов в исходной выборке, не попавших в нужную сторону от соответствующего коридора и степенью их удаленности от коридора.

1. Решающее дерево (Decision tree) -
https://learnmachinelearning.wikia.org/ru/wiki/Решающее_дерево_(Decision_tree)
2. Электронный учебник по статистике -
http://statsoft.ru/home/textbook/modules/stclatre.html

2. По какому принципу рассчитывается "важность признака (feature_importance)" в ансамблях деревьев?

Random Forest — алгоритм машинного обучения, предложенный Leo Breiman и Adele Cutler. Представляет собой ансамбль многочисленных, чувствительных к обучающей выборке алгоритмов (деревьев решений). Данные алгоритмы имеют маленькое смещение. Смещение (bias) метода обучения — это отклонение среднего ответа обученного алгоритма от ответа идеального алгоритма. Каждый из этих классификаторов строится на случайном подмножестве объектов и случайном подмножетсве признаков. Пусть обучающая выборка состоит из N примеров, размерность пространства признаков равна M, и задан параметр m. Запишем пошагово алгоритм Random Forest.

Все деревья ансамбля строятся независимо друг от друга по следующей процедуре:

1. Сгенерируем случайную подвыборку с повторением размером n из обучающей выборки.

2.Построим решающее дерево, классифицирующее примеры данной подвыборки, причём в ходе создания очередного узла дерева будем выбирать признак, на основе которого производится разбиение, не из всех M признаков, а лишь из m случайно выбранных.

3.Дерево строится до полного исчерпания подвыборки и не подвергается процедуре прунинга (англ. pruning — отсечение ветвей).

Классификация объектов проводится путём голосования: каждое дерево ансамбля относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

Алгоритм Random Forest может быть использован в задаче оценки важности признаков. Для этого необходимо обучить алгоритм на выборке и во время построения модели для каждого элемента обучающей выборки посчитать out-of-bag-ошибку. Пусть Xl n - бутстрапированная выборка дерева bn. Бутстрэппинг представляет собой выбор l объектов из выборки с возвращением, в результате чего некоторые объекты выбираются несколько раз, а некоторые — ни разу. Помещение нескольких копий одного объекта в бутстрапированную выборку соответствует выставлению веса при данном 4 объекте — соответствующее ему слагаемое несколько раз войдет в функционал, и поэтому штраф за ошибку на нем будет больше. Пусть L(y, z) — функция потерь, yi — ответ на i-м объекте обучающей выборки, тогда out-of-bag-ошибка вычисляется по следующей формуле:

$$\mathrm{OOB} = \sum_{i=1}^{l} L\left(y_i, \frac{1}{\sum_{n=1}^{N}\left[x_i \notin X_n^l\right]} \sum_{n=1}^{N}\left[x_i \notin X_n^l\right] b_n(x_i)\right)$$

Затем для каждого объекта такая ошибка усредняется по всему случайному лесу. Чтобы оценить важность признака, его значения перемешиваются для всех объектов обучающей выборки и out-of-bag-ошибка считается снова. Важность признака оценивается путем усреднения по всем деревьям разности показателей out-of-bag-ошибок до и после перемешивания значений. При этом значения таких ошибок нормализуются на стандартное отклонение. Случайный лес имеет еще некоторые преимущества для использования его в качестве алгоритма отбора признаков: он имеет очень мало настраиваемых параметров, относительно быстро и эффективно работает, что позволяет находить информативность признаков без значительных вычислительных затрат.

1. К. В. Воронцов Лекции по методам оценивания и выбора модели 2007

Библиотеки Python для Data Science: продолжение.

Урок 4. Оценка и интерпретация полученной модели. Обсуждение курсового проекта.