Машинное обучение в экономике

Семинар 3. Деревья

Задание №1

У вас имеется выборка из $n = 8$ наблюдений, характеризующих успешность стартапов в зависимости от наличия рекламной кампании и опытных членов команды, а также того, связан ли стартап с тематикой здоровья.

Успех	Реклама	Опыт	Здоровье
1	1	1	0
0	1	1	1
1	0	1	1
0	0	0	1
1	1	1	0
0	0	0	0
1	0	1	0
0	0	1	0

Вы обучаете решающее дерево глубины 2, прогнозирующее успех стартапа с помощью всех имеющихся в данных признаков (реклама, опыт и здоровье). В качестве критерия разбиения используется энтропия. Прогнозируется, что стартап окажется успешным, если условная вероятность этого события превышает 0.4.
Подсказка: вместо логарифма с основанием 2 эквивалентно использовать натуральный логарифм, необходимые значения которого указаны ниже:

$$\ln(1/8) \approx -2.079 \quad \ln(2/8) \approx -1.386 \quad \ln(3/8) \approx -0.981 \quad \ln(4/8) \approx -0.693$$
$$\ln(5/8) \approx -0.470 \quad \ln(6/8) \approx -0.288 \quad \ln(7/8) \approx -0.134 \quad \ln(8/8) = -0.000$$
$$\ln(1/6) \approx -1.792 \quad \ln(2/6) \approx -1.099 \quad \ln(4/6) \approx -0.405 \quad \ln(5/6) \approx -0.182$$
$$\ln(1/5) \approx -1.609 \quad \ln(2/5) \approx -0.916 \quad \ln(3/5) \approx -0.511 \quad \ln(4/5) = -0.223$$

1. Изобразите обученное решающее дерево графически, в каждом листе указав долю успешных стартапов.

2. Используя обученное решающее дерево спрогнозируйте, окажется ли успешным стартап с опытными участниками, без рекламы и посвященный тематике здоровья.

Задание №2

Вы прогнозируете вероятность дефолта по кредиту в зависимости от дохода индивида. Вы используете бэггинг, в котором в качестве базового используется метод **двух** ближайших соседей с расстоянием Манхэттен (для классификации).

Как в методе ближайших соседей, так и в ансамбле в случае равного количества 0 и 1 прогнозируется 1. Напомним, что в методе ближайших соседей в обучающей выборке наблюдение является одним из собственных ближайших соседей.

Доход_i	2	0	5	0	0	5	2	0	2	5	2	5	5	5	5
Дефолт_i	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0
Выборка	Исходная			Бутстрап 1			Бутстрап 2			Бутстрап 3			Бутстрап 4		

1. Получите прогноз дефолта для каждого наблюдения в исходной и бутстрапированной выборках. Результат представьте в форме таблицы.

2. Посчитайте OOB ошибку, руководствуясь критерием точности MAE.

Задание №3

Рассмотрим ансамбль из k решающих деревьев, основанный на бэггинге. Корреляция между прогнозами решающих деревьев равняется 0.6. Известно, что дисперсия прогноза ансамбля ровно в 1.25 раза меньше дисперсии прогноза одного решающего дерева.

1. Определите количество решающих деревье, используемых в ансамбле.

2. На тех же данных с использованием бэггинга оценили еще один аналогичный ансамбль с таким же количеством деревьев (значение k, найденное в предыдущем пункте). Найдите корреляцию между прогнозами этих двух ансамблей.

3. Определите, к чему будет стремиться корреляция между прогнозами ансамблей из предыдущих пунктов (с равным количеством деревьев) по мере стремления числа деревьев k к бесконечности. Сделайте вывод о том, насколько вероятно то, что при очень большом числе деревьев k эти ансамбли дадут существенно различающиеся прогнозы. Ответ подробно обоснуйте.

Задание №4

У вас имеется один бинарный признак X_i и целевая переменная Y_i, которая имеет условное (на признак) распределение Пуассона с параметром $\lambda = 2X_i$. Кроме того, известно, что $\mathrm{P}(X_i = 1) = 0.75$.

1. Найдите функцию $F(x)$, которая минимизирует среднеквадратическую ошибку прогноза целевой переменной.

2. Оцените условные на то, что признак принял значение x, смещение, дисперсию и шум прогнозов, получаемых с помощью функции, найденной в предыдущем пункте.

3. Допустим, что теперь для прогнозирования вы используете следующую функцию:

$$\hat{y}(x, X, Y) = x \times \frac{\sum_{i=1}^{n} X_i Y_i}{n}$$

Повторите предыдущий пункт для соответсвующего прогноза.