**Варианты лабораторной работы №3**

**по курсу «Машинное обучение. Часть 1»**

**Файл с данными**

data\_vN-NN.csv, N-NN – номер варианта

**Шаблон отчета**

<https://docs.google.com/presentation/d/1SaG1JYkH_X5eFrq6VleHDd757jmKna8wxb3YyGyK7h0/edit?usp=sharing>

**Отчет**

Файл otchet\_vN-NN\_GroupFIO.pdf и **исходные коды**, написанные для выполнения заданий, высылать на [mlearninglabs@gmail.com](mailto:mlearninglabs@gmail.com)

**Указания по оформлению графиков**

<http://datalearning.ru/study/Courses/methodic/lections/graphs.pdf>

К каждому графику в отчете должны быть даны все необходимые пояснения для его понимания (название, подписи осей, легенда, условия эксперимента, параметры обработки и пр.).

**Рекомендуемые средства программной реализации**

Python, MATLAB

**Литература**

1. Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

2. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

3. Raschka, S. (2015). *Python machine learning*. Packt Publishing Ltd.

**Вариант 3-01**

**Non-parametric one-dimensional density estimation**

**Задания**

1. Построить графики плотностей распределений, восстановленных с помощью наивного метода, для различных значений ширины окна: а) рассчитанной по правилу Сильвермана; б) слишком большой ширине окна; в) слишком малой ширине окна.

2. Построить графики плотностей распределений, восстановленных с использованием: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Ширину окна определить по правилу Сильвермана.

3. Построить графики зависимости дисперсии и смещения восстановленной плотности распределения от *x* для различных значений ширины гауссова окна: а) рассчитанной по правилу Сильвермана; б) слишком большой ширине окна; в) слишком малой ширине окна.

4. Построить графики зависимости дисперсии и смещения восстановленной плотности распределения при фиксированном *x* от ширины окна для: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна.

5. Построить графики зависимости средней интегральной ошибки восстановления плотности (MISE) от ширины окна для: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна.

6. Сделать выводы о влиянии вида и ширины окна на качество восстановления плотности распределения.

**Указания**

В пп.3, 4 для расчета дисперсии при фиксированном *x* восстанавливать плотность распределения по выборке 50 % случайно отобранных данных.

В пп.3, 4 при расчете смещения и в п.5 при расчёте MISE считать, что истинная функция плотности распределения имеет вид: , где  – функция плотности нормального распределения с мат. ожиданием *m* и с.к.о. σ.

Выводить значения MISE и дисперсий в логарифмическом масштабе.

**Литература**

1. Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

2. Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.

**Вариант 3-02**

**Non-parametric multi-dimensional density estimation**

**Задания**

1. Визуализировать плотность распределения, восстановленную с помощью наивного метода, для различных значений ширины окон: а) рассчитанной по правилу Сильвермана; б) слишком большой ширине окна; в) слишком малой ширине окна.

2. Визуализировать плотности распределения, восстановленные с использованием: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Ширину окон определить по правилу Сильвермана.

3. Визуализировать дисперсии и смещения восстановленной плотности распределения для каждой точки при различных значениях ширины гауссова окна: а) рассчитанной по правилу Сильвермана; б) слишком большой ширине окна; в) слишком малой ширине окна.

4. Построить графики зависимости дисперсии и смещения восстановленной плотности распределения в фиксированной точке (*x*1, *x*2) от коэффициента пропорциональности λ (отношение ширины парзеновского окна к ширине Сильвермана) для: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна.

5. Построить графики зависимости средней интегральной ошибки восстановления плотности (MISE) от коэффициента пропорциональности λ для а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна.

6. Сделать выводы о влиянии вида и ширины окна на качество восстановления плотности распределения.

**Указания**

Считать, что многомерные парзеновские ядра независимы.

Для визуализации двумерных функций плотности в пп.1,2, а также дисперсий и смещений в п.3 использовать поверхности (surface plots) и контурные диаграммы (contour plots).

В пп.3, 4 для расчета дисперсии в фиксированной точке (*x*1, *x*2) восстанавливать плотность распределения по выборке 50 % случайно отобранных данных.

В пп.3, 4 при расчете смещения и в п.5 при расчёте MISE считать, что истинная функция плотности распределения имеет вид: , где  – функция плотности двумерного нормального распределения с вектором мат. ожиданий *m* и ковариационной матрицей Σ, , , , ,, .

В п.4,5 вектор ширин парзеновских окон выбирать пропорциональным вектору ширины Сильвермана: .

Выводить значения MISE и дисперсий в логарифмическом масштабе.

**Литература**

1. Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.

**Вариант 3-03**

**Independent features vs independent kernels**

**Задания**

1. Визуализировать плотность распределения, восстановленную с помощью наивного метода, для различных значений ширины окон: а) рассчитанной по правилу Сильвермана; б) слишком большой ширине окна; в) слишком малой ширине окна.

2. В предположении о независимости признаков построить графики восстановленных одномерных плотностей распределения с использованием: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Ширину окна определить по правилу Сильвермана. Для каждого случая визуализировать соответствующую двумерную плотность.

3. В предположении о независимости парзеновских ядер визуализировать восстановленную двумерную плотность распределения с использованием: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Ширину окон определить по правилу Сильвермана.

4. Построить графики зависимости средней интегральной ошибки восстановления плотности (MISE) в предположениях о: а) независимости признаков (п.1); б) независимости ядер (п.2), от коэффициента пропорциональности λ (отношение ширины парзеновского окна к ширине Сильвермана) для окон а)–г).

5. Исследовать связь между MISE, рассчитанными в предположении о независимости признаков (MISE1) и в предположении о независимости ядер (MISE2). Построить диаграмму рассеяния в плоскости (MISE1, MISE2) (каждая точка соответствует некоторому значению λ). Рассчитать коэффициент корреляции, построить линейную регрессию.

6. Сделать выводы о качестве восстановления плотности распределения при предположениях о независимости признаков и независимости ядер.

**Указания**

Для визуализации двумерных функций плотности использовать поверхности (surface plots) и контурные диаграммы (contour plots).

В пп.4,5 при расчёте MISE считать, что истинная функция плотности распределения имеет вид: , где  – функция плотности двумерного нормального распределения с вектором мат. ожиданий *m* и ковариационной матрицей Σ, , , , ,, .

В п.4 вектор ширин парзеновских окон выбирать пропорциональным вектору ширины Сильвермана: .

Выводить значения MISE логарифмическом масштабе.

**Литература**

1. Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.

2. 1. Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

**Вариант 3-04**

**Binary logistic regression**

**Задания**

1. Обучить логистическую регрессионную модель с: а) logit-функцией связи; б) probit-функцией связи. Визуализировать области классов и границу между классами.

2. Рассчитать показатели точности обученных моделей: error rate, accuracy, sensitivity, specificity, precision, fall-out, F1-score, Cohen’s kappa на обучающей и тестовой выборках.

3. Построить ROC-кривые и PR-кривые по результатам классификации, рассчитать ROC AUC и PR AUC.

4. Обучить нормальный наивный байесовский классификатор в предположении о равных ковариационных матрицах классов. Записать аналитическое выражение для границы классов. Визуализировать области классов и границу между классами, сравнить с границами классов, полученными в п.1.

5. Построить графики зависимости точности классификации на обучающей и тестовой выборках от параметра регуляризации для обученной регуляризованной логистической регрессии с: а) L2-регуляризатором; б) L1-регуляризатором. Визуализировать границу между классами для нескольких различных значений параметра регуляризации, сравнить с границей, полученной в п.1.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Кросс-валидация: Holdout (70/30).

В п.4 априорные вероятности классов считать равными.

**Литература**

1. Hilbe, J. M. (2016). *Practical guide to logistic regression*. crc Press.

2. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

3. Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

**Вариант 3-05**

**Multiclass logistic regression**

**Задания**

1. Обучить многоклассовую логистическую регрессионную модель, визуализировать полученные области классов и границы между классами.

2. Обучить бинарные логистические регрессионные модели для каждого класса (с использованием OVA-кодирования). Визуализировать границы между классами, формируемые каждой моделью. Рассчитать micro-averaged и macro-averaged показатели качества: error rate, accuracy, sensitivity, specificity, precision, fall-out, F1-score, Cohen’s kappa на обучающей и тестовой выборках.

3. Построить ROC-кривые и PR-кривые для каждого класса, micro-averaged и macro-averaged ROC-кривые и PR-кривые, рассчитать ROC AUC и PR AUC для каждого класса, рассчитать micro-averaged и macro-averaged ROC AUC и PR AUC.

4. Обучить нормальный наивный байесовский классификатор в предположении о равных ковариационных матрицах классов. Записать аналитические выражения для границ классов. Визуализировать области классов и границы между классами, сравнить с границами классов, полученными в п.1.

5. Построить графики зависимости точности классификации на обучающей и тестовой выборках от параметра регуляризации для обученной регуляризованной логистической регрессии с: а) L2-регуляризатором; б) L1-регуляризатором. Визуализировать границу между классами для нескольких различных значений параметра регуляризации, сравнить с границей, полученной в п.1.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Кросс-валидация: Holdout (70/30).

В п.4 априорные вероятности классов считать равными.

**Литература**

1. Hilbe, J. M. (2016). *Practical guide to logistic regression*. crc Press.

2. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

3. Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

**Вариант 3-06**

**Kernel regression**

**Задания**

1. Построить регрессограммы для различного числа интервалов группировки: а) 3; б) 10; в) 50. Построить график зависимости MSE непараметрической регрессии на обучающей, валидационной и тестовой выборках от числа интервалов группировки.

2. Построить графики зависимости MSE ядерной регрессии на обучающей, валидационной и тестовой выборках от ширины окна для: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Определить оптимальную ширину каждого окна.

3. Визуализировать графики ядерной функции регрессии, полученной с использованием: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Ширину окна задать: i) равной соответствующему оптимальному значению, полученному в п.2; ii) значению, меньшему оптимального; iii) значению, большему оптимального.

4. Построить графики зависимости дисперсии и смещения регрессионной модели для нескольких фиксированных значений *x* от ширины окна для: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна.

5. Построить график зависимости MSE KNN-регрессии от числа ближайших соседей K на обучающей и тестовой выборках. Определить оптимальное число ближайших соседей и визуализировать графики KNN-регрессии для различных значений K: а) равному оптимальному; б) меньшего оптимального; в) большего оптимального.

6. Сделать выводы о влиянии вида и ширины окна на качество ядерной регрессионной модели.

**Указания**

Разбиение на обучающую, валидационную и тестовую выборки – holdout в отношении 60/30/10.

В пп.2, 5 оптимальные значения гиперпараметров обучения определять по значению ошибки на валидационной выборке.

**Литература**

1. Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.

2. Fan, J., and I. Gijbels. (1996). *Local Polynomial Modelling and Its Applications*. Vol. 66. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

3. Chu, C. K., & Marron, J. S. (1991). Choosing a kernel regression estimator. *Statistical science*, 404-419.

4. Wand, M. P., and M. C. Jones. (1995). *Kernel Smoothing*. Vol. 60. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

**Вариант 3-07**

**LOESS estimator**

**Задания**

1. Построить графики зависимости MSE локальной линейной регрессии на обучающей, валидационной и тестовой выборках от ширины окна для: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Определить оптимальную ширину каждого окна.

2. Визуализировать графики локальной линейной регрессии, полученной с использованием: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Ширину окна задать: i) равной соответствующему оптимальному значению, полученному в п.1; ii) значению, меньшему оптимального; iii) значению, большему оптимального.

3. Построить графики зависимости MSE LOESS-регрессии на обучающей, валидационной и тестовой выборках от степени локального полинома для: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Ширину окна задать равной соответствующему оптимальному значению, полученному в п.1. Визуализировать графики LOESS-регрессии для оптимальной степени локального полинома.

4. Построить графики зависимости дисперсии и смещения локальной линейной регрессии для нескольких фиксированных значений *x* от ширины окна для: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна.

5. Построить графики зависимости MSE полиномиальной регрессии на обучающей, валидационной и тестовой выборках от степени полинома. Сравнить оптимальную полиномиальную функцию регрессии с оптимальной локальной линейной функцией регрессии.

6. Сделать выводы о влиянии параметров LOESS-регрессии на ее качество.

**Указания**

Разбиение на обучающую, валидационную и тестовую выборки – holdout в отношении 60/30/10.

В пп.1, 5 оптимальные значения гиперпараметров обучения определять по значению ошибки на валидационной выборке.

**Литература**

1. Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.

2. Takezawa, K. (2005). *Introduction to nonparametric regression*. John Wiley & Sons.

3. Jacoby, W. G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral studies*, *19*(4), 577-613.

**Вариант 3-08**

**Regression trees (single feature)**

**Задания**

1. Построить график зависимости MSE обученного регрессионного дерева на обучающей, валидационной и тестовой выборках от: а) глубины дерева; б) числа ветвлений дерева. Определить оптимальное число ветвлений, визуализировать оптимальное дерево и графики регрессионных остатков для них. Для нескольких фиксированных значений глубины дерева построить графики зависимостей MSE откликов в вершинах дерева от номера вершины.

2. Построить графики зависимостей: а) MSE обученной регрессионной модели на обучающей и валидационной выборках; б) глубины; в) количества ветвлений обученного дерева, от минимального числа обучающих примеров в терминальных вершинах дерева.

3. Построить график зависимости MSE обученного дерева на обучающей и валидационной выборках от объема обучающей выборки для нескольких фиксированных значений: а) глубины; б) числа ветвлений регрессионного дерева (в т.ч. для оптимального, полученного в п.1).

4. Обучить регрессионное дерево с избыточным числом вершин и провести его обрезку (pruning). Построить графики зависимости MSE на обучающей, валидационной и тестовой выборках от числа вершин дерева после обрезки. Проводить обрезку с использованием стратегий: а) reduced error pruning; б) minimum cost complexity pruning.

5. Для каждого региона, формируемого оптимальным деревом, полученным в п.1, построить простейшую линейную регрессионную модель. Визуализировать остатки полученной кусочно-линейной регрессионной модели, сравнить MSE полученной модели и оптимального регрессионного дерева на обучающей, валидационной и тестовой выборках.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Разбиение на обучающую, валидационную и тестовую выборки – holdout в отношении 60/30/10.

Оптимальные значения гиперпараметров обучения определять по значению ошибки на валидационной выборке.

При обучении деревьев в п.1 не ограничивать минимальное число обучающих примеров в терминальных вершинах.

**Литература**

1. Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

2. Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

**Вариант 3-09**

**Regression trees (multiple features)**

**Задания**

1. Построить зависимости MSE обученного регрессионного дерева на обучающей, валидационной и тестовой выборках от максимального числа ветвлений дерева и минимального числа обучающих примеров в терминальных вершинах дерева. Визуализировать построенные зависимости в виде heatmap или поверхности. Определить оптимальные значения указанных гиперпараметров.

2. Построить графики зависимости MSE обученного регрессионного дерева на обучающей, валидационной и тестовой выборках от количества признаков, рассматриваемых в каждой вершине перед ветвлением для: а) гиперпараметров обучения, равных полученным в п.1; б) без ограничений на число ветвлений дерева и минимальное число обучающих примеров в терминальных вершинах дерева. Определить оптимальное значение количества признаков, рассматриваемых в каждой вершине перед ветвлением.

3. Обучить регрессионное дерево с избыточным числом вершин и провести его обрезку (pruning). Построить графики зависимости MSE на обучающей, валидационной и тестовой выборках от числа вершин дерева после обрезки. Проводить обрезку с использованием стратегий: а) reduced error pruning; б) minimum cost complexity pruning.

4. Для каждой вершины найденных в пп.1-3 оптимальных деревьев рассчитать выигрыш в значении MSE после ветвления. На основе полученных значений для каждого признака рассчитать важность, используя: а) невзвешенный; б) взвешенный (объемами выборки в вершинах) способы расчета. Сделать вывод о важности признаков, используемых в регрессионном дереве.

5. Построить графики зависимости значений коэффициентов линейной L1-регуляризованной регрессионной модели от параметра LASSO-регрессии (LASSO trace plot). Сделать вывод о важности признаков, используемых в линейной регрессионной модели.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Разбиение на обучающую, валидационную и тестовую выборки – holdout в отношении 60/30/10.

При обучении деревьев в п.1 использовать все признаки.

В п.2 провести обучение несколько раз, выводить на график среднее и с.к.о. значений MSE (по обучениям).

Оптимальные значения гиперпараметров обучения определять по значению ошибки на валидационной выборке.

**Литература**

1. Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

2. Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

**Вариант 3-10**

**Classification trees (single feature)**

**Задания**

1. Обучить классификационные деревья с использованием для оценивания качества разбиения узлов метрик: а) индекс Джини; б) энтропия распределения меток классов; в) ошибка классификации. Для каждого дерева построить графики зависимостей метрик от номера узла.

2. Построить графики зависимости ошибки классификации для обученного дерева на обучающей, валидационной и тестовой выборках от: а) глубины дерева; б) числа ветвлений дерева, при использовании в обучении метрик: а) индекс Джини; б) энтропия распределения меток классов; в) ошибка классификации. Для каждой метрики определить оптимальное число ветвлений дерева, визуализировать оптимальное дерево и формируемые им границы классов.

3. Построить графики зависимости ошибки классификации для обученного дерева на обучающей, валидационной и тестовой выборках от минимального числа обучающих примеров в терминальных вершинах дерева при использовании в обучении метрик: а) индекс Джини; б) энтропия распределения меток классов; в) ошибка классификации. Для каждой метрики визуализировать оптимальное дерево и формируемые им границы классов.

4. Провести обрезку (pruning) деревьев, построенных в п.1. Построить графики зависимости ошибки классификации на обучающей, валидационной и тестовой выборках от числа вершин дерева после обрезки. Проводить обрезку с использованием стратегий: а) reduced error pruning; б) minimum cost complexity pruning. Для оптимальных деревьев визуализировать формируемые границы классов.

5. Обучить многоклассовую логистическую регрессионную модель, визуализировать полученные области классов и границы между классами. Сравнить ошибку классификации и границы классов, формируемые классификационным деревом и логистической регрессионной моделью.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Разбиение на обучающую, валидационную и тестовую выборки – holdout в отношении 60/30/10.

Оптимальные значения гиперпараметров обучения определять по значению ошибки на валидационной выборке.

При обучении деревьев в п.1 не ограничивать число вершин дерева и минимальное число обучающих примеров в терминальных вершинах.

**Литература**

1. Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

2. Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

**Вариант 3-11**

**Classification trees (multiple features)**

**Задания**

1. Построить зависимости ошибки классификации обученного регрессионного дерева на обучающей, валидационной и тестовой выборках от максимального числа ветвлений дерева и минимального числа обучающих примеров в терминальных вершинах дерева при использовании метрик оценивания качества разбиения узлов: а) индекс Джини; б) энтропия распределения меток классов; в) ошибка классификации. Визуализировать построенные зависимости в виде heatmap или поверхности. Определить оптимальные значения указанных гиперпараметров.

2. Построить графики зависимости ошибки классификации обученного классификационного дерева на обучающей, валидационной и тестовой выборках от количества признаков, рассматриваемых в каждой вершине перед ветвлением для: а) гиперпараметров обучения, равных полученным в п.1; б) без ограничений на число ветвлений дерева и минимальное число обучающих примеров в терминальных вершинах дерева. Определить оптимальное значение количества признаков, рассматриваемых в каждой вершине перед ветвлением.

3. Обучить классификационное дерево с избыточным числом вершин и провести его обрезку (pruning). Построить графики зависимости ошибки классификации на обучающей, валидационной и тестовой выборках от числа вершин дерева после обрезки. Проводить обрезку с использованием стратегий: а) reduced error pruning; б) minimum cost complexity pruning.

4. Для каждой вершины найденных в пп.1-3 оптимальных деревьев рассчитать выигрыш в значениях метрик после ветвления: а) индекса Джини; б) энтропии распределения меток классов; в) ошибки классификации. На основе полученных значений для каждого признака рассчитать важность, используя: а) невзвешенный; б) взвешенный (объемами выборки в вершинах) способы расчета. Сделать вывод о важности признаков, используемых в классификационном дереве.

5. Обучить многоклассовую логистическую регрессионную модель. Сравнить ошибку классификации классификационного дерева и логистической регрессионной модели. Оценить важность признаков для логистической регрессионной модели, используя метод пермутаций.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Разбиение на обучающую, валидационную и тестовую выборки – holdout в отношении 60/30/10.

При обучении деревьев в п.1 использовать все признаки.

В п.2 провести обучение несколько раз, выводить на график среднее и с.к.о. ошибок классификации (по обучениям).

Оптимальные значения гиперпараметров обучения определять по значению ошибки на валидационной выборке.

**Литература**

1. Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

2. Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

**Вариант 3-12**

**Non-parametric regression techniques**

**Задания**

1. Построить графики зависимости MSE обученной многомерной ядерной регрессионной модели с независимыми ядрами на обучающей, валидационной и тестовой выборках от ширины окна для: а) прямоугольного окна; б) гауссова окна. Определить оптимальную ширину каждого окна.

2. Построить графики зависимости MSE KNN-регрессии от числа ближайших соседей K на обучающей, валидационной и тестовой выборках. Определить оптимальное число ближайших соседей.

3. Построить графики зависимости MSE локальной линейной регрессии LOESS (с независимыми ядрами) на обучающей, валидационной и тестовой выборках от ширины окна для: а) прямоугольного окна; б) гауссова окна. Определить оптимальную ширину каждого окна.

4. Построить графики зависимости MSE регрессионного дерева от числа ветвлений дерева на обучающей, валидационной и тестовой выборках. Определить оптимальное число ветвлений дерева.

5. Обучить множественную линейную регрессионную модель, используя метод OLS, Рассчитать MSE обученной модели на обучающей и тестовой выборках.

6. Сделать выводы по результатам сравнения обученных регрессионных моделей.

**Указания**

Разбиение на обучающую, валидационную и тестовую выборки – holdout в отношении 60/30/10.

Оптимальные значения гиперпараметров определять по значению ошибки на валидационной выборке.

Гиперпараметры, используемые при обучении регрессионного дерева в п.4, подобрать экспериментально.

**Литература**

1. Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.

2. Takezawa, K. (2005). *Introduction to nonparametric regression*. John Wiley & Sons.

3. Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

**Вариант 3-13**

**Non-parametric classification techniques**

**Задания**

1. В предположении о независимости парзеновских ядер визуализировать восстановленные двумерные плотности распределения для каждого класса для случаев: а) прямоугольного окна; б) гауссова окна. Ширину окон определить по правилу Сильвермана.

2. Построить графики зависимости ошибки обученного непараметрического байесовского классификатора на обучающей, валидационной и тестовой выборках от коэффициента пропорциональности λ (отношение ширины парзеновского окна к ширине Сильвермана) для случаев: а) прямоугольного окна; б) гауссова окна. Для каждого типа окна определить коэффициент λ, при котором байесовский классификатор обладает наибольшей обобщающей способностью. Сравнить восстановленные плотности распределения с полученными в п.1. Для оптимальных классификаторов визуализировать формируемые области классов и границы между классами.

3. Построить графики зависимости ошибки обученного KNN-классификатора от числа ближайших соседей K на обучающей, валидационной и тестовой выборках. Определить оптимальное число ближайших соседей. Для оптимального классификатора визуализировать формируемые области классов и границы между классами.

4. Построить графики зависимости ошибки классификационного дерева от числа ветвлений дерева на обучающей, валидационной и тестовой выборках. Определить оптимальное число ветвлений дерева. Для оптимального дерева визуализировать формируемые области классов и границы между классами.

5. Обучить многоклассовую логистическую регрессионную модель, визуализировать полученные области классов и границы между классами.

6. Сделать выводы по результатам сравнения обученных классификаторв и формируемых ими границ классов.

**Указания**

Разбиение на обучающую, валидационную и тестовую выборки – holdout в отношении 60/30/10.

Оптимальные значения гиперпараметров определять по значению ошибки на валидационной выборке.

В п.2 априорные вероятности классов считать равными; вектор ширин парзеновских окон выбирать пропорциональным вектору ширины Сильвермана: .

Гиперпараметры, используемые при обучении регрессионного дерева в п.4, подобрать экспериментально.

**Литература**

1. Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.

3. Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

**Вариант 1-14**

**Bagged regression trees**

**Задания**

1. Построить график зависимости MSE обученного регрессионного дерева на обучающей, валидационной и тестовой выборках от: а) глубины дерева; б) числа ветвлений дерева. Определить оптимальное число ветвлений, визуализировать оптимальное дерево, обученную функцию регрессии и регрессионные остатки.

2. Построить зависимости MSE ансамбля на обучающей и тестовой выборках, а также OOB-ошибки ансамбля от объема бутстреп-выборок, используемых для обучения отдельных деревьев, и числа деревьев в ансамбле. Визуализировать полученную зависимость в виде heatmap, поверхности или контурной диаграммы.

3. Рассчитать коэффициенты корреляции выходов, формируемых отдельными деревьями ансамбля. Предложить показатель, характеризующий коррелированность выходов всех деревьев ансамбля. Построить зависимость этого показателя от объема бутстреп-выборок, используемых для обучения отдельных деревьев, и числа деревьев в ансамбле. Визуализировать полученную зависимость в виде heatmap, поверхности или контурной диаграммы.

4. Исследовать зависимость обобщающих способностей ансамбля от сложности входящих в него деревьев. Для ансамбля с фиксированным числом деревьев *K* построить график зависимости OOB-ошибки ансамбля от: а) максимального числа ветвлений деревьев ансамбля; б) минимального числа обучающих примеров в терминальных вершинах деревьев (все остальные гиперпараметры обучения деревьев зафиксировать). Повторить для нескольких различных значений *K*.

5. Построить зависимости OOB-ошибки и ошибки *k*-fold кросс-валидации ансамбля от числа деревьев в ансамбле при использовании для обучения деревьев бутстреп-выборок различного объема. Исследовать статистическую связь между OOB-ошибкой и ошибкой кросс-валидации ансамбля при различных значениях параметра *k* кросс-валидации.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Оптимальные значения гиперпараметров обучения в п.1 определять по значению ошибки на валидационной выборке (при holdout-разбиении на обучающую, валидационную и тестовую выборки – в отношении 60/30/10).

При обучении деревьев в п.1 не ограничивать минимальное число обучающих примеров в терминальных вершинах.

**Литература**

1. Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

2. Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

**Вариант 3-15**

**Random forest regression**

**Задания**

1. Построить график зависимости MSE обученного регрессионного дерева на обучающей, валидационной и тестовой выборках от: а) глубины дерева; б) числа ветвлений дерева. Определить оптимальное число ветвлений, визуализировать оптимальное дерево и графики регрессионных остатков для них. Для нескольких фиксированных значений глубины дерева построить графики зависимостей MSE откликов в вершинах дерева от номера вершины.

2. Построить графики зависимостей MSE ансамбля на обучающей и тестовой выборках от числа деревьев в ансамбле при использовании для обучения деревьев бутстреп-выборок объема, равного: а) объему обучающей выборки; б) половине объема обучающей выборки, и числа признаков в вершинах, равного: i) исходному числу признаков; ii) половине исходного числа признаков.

3. Для ансамбля с фиксированным числом деревьев *K* построить зависимость MSE ансамбля на обучающей и тестовой выборках, а также OOB-ошибки ансамбля от объема бутстреп-выборок, используемых для обучения отдельных деревьев, и числа признаков в вершинах деревьев. Визуализировать полученную зависимость в виде heatmap, поверхности или контурной диаграммы. Повторить для нескольких различных значений *K*.

4. Для ансамбля с фиксированным числом деревьев *K* рассчитать коэффициенты корреляции выходов, формируемых отдельными деревьями ансамбля. Предложить показатель, характеризующий коррелированность выходов всех деревьев ансамбля. Построить зависимость этого показателя от объема бутстреп-выборок, используемых для обучения отдельных деревьев, и числа признаков в вершинах деревьев. Визуализировать полученную зависимость в виде heatmap, поверхности или контурной диаграммы. Повторить для нескольких различных значений *K*.

5. Построить графики зависимостей OOB-ошибки и ошибки 5-fold кросс-валидации ансамбля от числа деревьев в ансамбле при использовании для обучения деревьев бутстреп-выборок объема, равного: а) объему обучающей выборки; б) половине объема обучающей выборки, и числа признаков в вершинах, равного: i) исходному числу признаков; ii) половине исходного числа признаков. Исследовать статистическую связь между OOB-ошибкой и ошибкой кросс-валидации ансамбля.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Оптимальные значения гиперпараметров обучения в п.1 определять по значению ошибки на валидационной выборке (при holdout-разбиении на обучающую, валидационную и тестовую выборки – в отношении 60/30/10).

При обучении деревьев в п.1 не ограничивать минимальное число обучающих примеров в терминальных вершинах.

При построении ансамблей гиперпараметры обучения отдельных деревьев зафиксировать и положить равными полученным в п.1.

**Литература**

1. Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

2. Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

**Вариант 3-16**

**Random forest classification**

**Задания**

1. Построить зависимости ошибки классификации обученного классификационного дерева на обучающей, валидационной и тестовой выборках от максимального числа ветвлений дерева и минимального числа обучающих примеров в терминальных вершинах дерева при использовании метрик оценивания качества разбиения узлов: а) индекс Джини; б) энтропия распределения меток классов; в) ошибка классификации. Визуализировать построенные зависимости в виде heatmap, поверхности или контурной диаграммы. Определить оптимальные значения указанных гиперпараметров.

2. Построить графики зависимостей ошибки классификации обученного ансамбля на обучающей и тестовой выборках от числа деревьев в ансамбле при использовании для обучения деревьев бутстреп-выборок объема, равного: а) объему обучающей выборки; б) половине объема обучающей выборки, и числа признаков в вершинах, равного: i) исходному числу признаков; ii) половине исходного числа признаков (всего 4 графика).

3. Для ансамбля с фиксированным числом деревьев *K* построить зависимость ошибки классификации обученного ансамбля на обучающей и тестовой выборках, а также OOB-ошибки ансамбля от объема бутстреп-выборок, используемых для обучения отдельных деревьев, и числа признаков в вершинах деревьев. Визуализировать полученную зависимость в виде heatmap, поверхности или контурной диаграммы. Повторить для нескольких различных значений *K*.

4. Исследовать зависимость обобщающих способностей ансамбля от сложности входящих в него деревьев. Для ансамбля с фиксированным числом деревьев *K* построить график зависимости OOB-ошибки ансамбля от: а) максимального числа ветвлений деревьев ансамбля; б) минимального числа обучающих примеров в терминальных вершинах деревьев (все остальные гиперпараметры обучения деревьев зафиксировать). Повторить для нескольких различных значений *K*.

5. Исследовать зависимость обобщающих способностей ансамбля от объема обучающей выборки. Построить графики зависимостей ошибки классификации обученного ансамбля на обучающей и тестовой выборках, а также OOB-ошибки ансамбля от объема обучающей выборки для некоторых фиксированных значений числа деревьев в ансамбле, объемов бутстреп-выборок, изпользуемых для обучения деревьев, и числа признаков в вершинах деревьев.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Оптимальные значения гиперпараметров обучения в п.1 определять по значению ошибки на валидационной выборке (при holdout-разбиении на обучающую, валидационную и тестовую выборки – в отношении 60/30/10).

При обучении деревьев в п.1 использовать все признаки.

При построении ансамблей гиперпараметры обучения отдельных деревьев зафиксировать и положить равными полученным в п.1.

**Литература**

1. McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41-48).

2. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.