**Варианты лабораторной работы №2**

**по курсу «Анализ данных и машинное обучение»**

**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 2-01**

**Binary classification loss functions and empirical risk**

**Задания**

1. Рассчитать отступы классификатора для каждого примера. Построить график зависимости отступа от номера примера (номера отсортировать в порядке возрастания отступов).

2. Рассчитать значения функции потерь по заданным результатам классификации: 0-1 loss, hinge loss, logistic loss для каждого примера. Построить графики зависимости значений функций потерь от номера примера. Указать на графике правильно классифицированные и неправильно классифицированные примеры.

3. Построить гистограммы распределений значений различных функций потерь.

4. Построить диаграмму рассеяния исходных признаков (примеры разных классов отметить разными цветами). Отметить на диаграмме примеры, имеющие высокие значения функции потерь (для 0-1 loss, hinge loss, logistic loss).

5. Построить графики зависимости эмпирического риска при различных функциях потерь от значения порога классификатора.

6. Сделать выводы о результатах классификации, сравнить различные функции потерь и соответствующие значения эмпирического риска.

**Указания**

При построении графиков п.2 отсортировать номера примеров в порядке убывания значений логистической функции потерь.

**Литература**

1. Rosasco, L., De Vito, E., Caponnetto, A., Piana, M., & Verri, A. (2004). Are loss functions all the same?. *Neural Computation*, *16*(5), 1063-1076.

2. Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, second edition. Springer, New York, 2008.

**Вариант 2-02**

**Binary confusion matrix based measures**

**Задания**

1. Построить таблицу сопряженности по заданным результатам бинарной классификации.

2. Рассчитать показатели, основанные на таблице сопряженности: error rate, accuracy, sensitivity, specificity, precision, fall-out, F1-score, Cohen’s kappa.

3. Построить график зависимости Fβ(β).

4. Построить графики зависимости чувствительности и специфичности от объёма выборки.

5. Сравнить значения показателей, основанных на таблице сопряженности, для случая, если положительные и отрицательные классы инвертированы (т.е. если все метки +1 заменить на -1, и все метки -1 – на +1).

6. Сделать выводы о качестве классификации на основе рассчитанных значений показателей.

**Указания**

В п.4 подвыборки формировать случайным образом, сохраняя пропорции классов.

**Литература**

1. Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27-38.

2. Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, *217*(1), 48-58.

**Вариант 2-03**

**Binary model-wide measures**

**Задания**

1. Построить ROC-кривую и PR-кривую по заданным результатам классификации, рассчитать ROC AUC и PR AUC.

2. Проверить гипотезу о превышении классификационных очков положительного класса над очками отрицательного класса, используя критерий Манна-Уитни (привести значения статистики критерия, *p-value*). Сравнить нормированное значение статистики Манна-Уитни со значением показателя ROC AUC.

3. Отметить точки на ROC-кривой и PR-кривой, соответствующие оптимальным значениям порога (по различным критериям: max Youden’s index, max informedness и пр.).

4. Построить графики зависимости ROC AUC и PR AUC от объёма выборки.

5. По данным п.4 построить диаграмму рассеяния значений ROC AUC и PR AUC, рассчитать коэффициент корреляции между этими показателями.

6. Сделать выводы об использовании ROC-кривой, PR-кривой и показателей ROC AUC и PR AUC для оценки качества классификации.

**Указания**

В п.4 подвыборки формировать случайным образом, сохраняя пропорции классов.

**Литература**

1. Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240). ACM.

2. Cortes, C., & Mohri, M. (2004). AUC optimization vs. error rate minimization. In *Advances in neural information processing systems* (pp. 313-320).

3. Lobo, J. M., Jiménez‐Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, *17*(2), 145-151.

4. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

5. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

**Вариант 2-04**

**Multiclass confusion matrix based measures**

**Задания**

1. Построить таблицу сопряженности по заданным результатам многоклассовой классификации.

2. Построить таблицы сопряженности бинарной OVA-классификации для каждого класса.

3. Рассчитать micro-averaged и macro-averaged показатели, основанные на таблице сопряженности: error rate, accuracy, sensitivity, specificity, precision, fall-out, F1-score, Cohen’s kappa.

4. Построить графики зависимости Fβ(β) для каждого класса и macro-averaged Fβ(β).

5. Построить графики зависимости чувствительности и специфичности для каждого класса, а также micro-averaged и macro-averaged чувствительности и специфичности, от объёма выборки.

6. Сделать выводы о показателях качества многоклассовой классификации, основанных на таблице сопряженности.

**Указания**

В п.5 подвыборки формировать случайным образом, сохраняя пропорции классов.

**Литература**

1. Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27-38.

**Вариант 2-05**

**Multiclass model-wide measures**

**Задания**

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

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

3. Построить графики зависимости micro-averaged и macro-averaged ROC AUC и PR AUC от объёма выборки.

4. По данным п.3 проверить гипотезу об однородности значений micro-averaged и macro-averaged ROC AUC, используя критерий Манна-Уитни (привести значения статистики критерия, *p-value*).

5. Исключить из выборки примеры из наибольшего (по числу примеров) класса и рассчитать micro-averaged и macro-averaged ROC AUC и PR AUC. Сравнить полученные значения с аналогичными для исходной выборки.

6. Сделать выводы об использовании micro-averaged и macro-averaged ROC-кривых, PR-кривых и показателей ROC AUC и PR AUC для оценки качества многоклассовой классификации.

**Указания**

В п.3 подвыборки формировать случайным образом, сохраняя пропорции классов.

**Литература**

1. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437.

**Вариант 2-06**

**ECOC coding design schemes**

**Задания**

1. Обучить байесовские бинарные классификаторы при различных способах ECOC-кодирования: а) OVA; б) OVO; в) полное бинарное кодирование; г) полное тернарное кодирование.

2. Для каждого из полученных наборов базовых классификаторов рассчитать классификационные очки многоклассового классификатора, используя различные способы ECOC-декодирования: а) взвешенное; б) невзвешенное.

3. В исходном пространстве признаков изобразить области классов (закрасить разными цветами), формируемые каждым из полученных в п.2 многоклассовым классификатором (всего 8 классификаторов). Нанести на диаграммы границы классов и данные из обучающей и тестовой выборок. На отдельной диаграмме изобразить все границы классов, формируемые построенными многоклассовыми классификаторами.

4. Рассчитать точности (accuracy) построенных многоклассовых классификаторов на обучающей и тестовой выборках, представить данные в графическом виде (bar charts).

5. Построить micro-averaged и macro-averaged ROC-кривые и PR-кривые и рассчитать ROC AUC и PR AUC для всех построенных многоклассовых классификаторов.

6. Сделать вывод о влиянии способов ECOC-кодирования и ECOC-декодирования на точность многоклассовой классификации.

**Указания**

Кросс-валидация: Holdout (70/30).

При обучении бинарных байесовских классификаторов считать, что обучающие выборки для каждого класса получены из нормального распределения (без предположений о ковариационных матрицах).

Априорные вероятности классов для каждого бинарного классификатора считать равными.

Классификационные очки бинарных классификаторов считать по формуле , где *pk* – апостериорная вероятность *k*-го класса.

**Литература**

1. Escalera, S., Pujol, O., & Radeva, P. (2009). Separability of ternary codes for sparse designs of error-correcting output codes. *Pattern Recognition Letters*, *30*(3), 285-297.

2. Escalera, S., Pujol, O., & Radeva, P. (2010). Re-coding ECOCs without re-training. *Pattern Recognition Letters*, *31*(7), 555-562.

**Вариант 2-07**

**ECOC random coding design**

**Задания**

1. Обучить байесовские бинарные классификаторы при случайном ECOC-кодировании: а) бинарном (dense); б) тернарном (sparse). Число бинарных классификаторов, участвующих в кодировании, задать произвольно.

2. Для каждого из полученных наборов базовых классификаторов рассчитать классификационные очки многоклассового классификатора, используя различные способы ECOC-декодирования: а) взвешенное; б) невзвешенное.

3. Построить графики зависимости точности (accuracy) полученных в п.2 многоклассовых классификаторов на обучающей и тестовой выборках от числа бинарных классификаторов, участвовавших в случайном ECOC-кодировании.

4. Для полученных в п.2 многоклассовых классификаторов построить графики зависимостей micro-averaged и macro-averaged ROC AUC и PR AUC на обучающей и тестовой выборках от числа бинарных классификаторов, участвовавших в случайном ECOC-кодировании.

5. В исходном пространстве признаков изобразить области классов (закрасить разными цветами), формируемые многоклассовым классификатором, обладающим наибольшей точностью (accuracy), при а) взвешенном; б) невзвешенном ECOC-декодировании. Нанести на диаграммы границы классов и данные из обучающей и тестовой выборок.

6. Сделать вывод о влиянии числа бинарных классификаторов, участвующих в случайном ECOC-кодировании, на точность классификации.

**Указания**

Кросс-валидация: Holdout (70/30).

При обучении бинарных байесовских классификаторов считать, что обучающие выборки для каждого класса получены из нормальных распределений (без предположений о ковариационных матрицах).

Априорные вероятности классов для каждого бинарного классификатора считать равными.

Классификационные очки бинарных классификаторов считать по формуле , где *pk* – апостериорная вероятность *k*-го класса.

Столбцы матрицы случайного бинарного (тернарного) кодирования отбирать случайным образом (без повторений) из столбцов матрицы полного бинарного (тернарного) кодирования,

**Литература**

1. Escalera, S., Pujol, O., & Radeva, P. (2009). Separability of ternary codes for sparse designs of error-correcting output codes. *Pattern Recognition Letters*, *30*(3), 285-297.

2. Escalera, S., Pujol, O., & Radeva, P. (2010). Re-coding ECOCs without re-training. *Pattern Recognition Letters*, *31*(7), 555-562.

**Вариант 2-08**

**Mahalanobis distance-based classifier**

**Задания**

1. Рассчитать расстояния Махаланобиса от каждой точки до каждого из классов при различных предположениях: а) ковариационные матрицы всех классов равные, скалярные; б) равные, диагональные; в) различные, скалярные; г) различные, диагональные; д) равные; е) различные. Визуализировать матрицы для каждого случая. 2. Построить диаграммы рассеяния расстояний (по оси абсцисс ρ1 – расстояние до положительного класса (с меткой +1), по оси ординат ρ2 – до отрицательного класса) для каждого случая. Нанести на диаграмму прямую ρ1 = ρ2. Построить гистограммы распределения расстояний данных каждого класса до положительного класса для случаев а)–е).

3. На диаграмме рассеяния исходных данных изобразить границы классов (геометрическое место точек, равноудалённых по Махаланобису от обоих классов). Сравнить границы для случаев а)–е).

4. Рассчитать показатели точности (error rate, accuracy, sensitivity, specificity, precision, fall-out, F1-score, Cohen’s kappa) на обучающей и тестовой выборках для построенных классификаторов (для случаев а)–е)).

5. Построить ROC-кривые и PR-кривые и рассчитать ROC AUC и PR AUC на обучающей и тестовой выборках для случаев а)–е).

6. Сделать выводы о влиянии предположений о модели обучающих данных на точность классификации по расстоянию Махаланобиса.

**Указания**

Кросс-валидация: Holdout (70/30).

В качестве классификационных очков при построении ROC-кривых и PR-кривых использовать разность между расстояниями Махаланобиса до положительного и отрицательного классов.

**Литература**

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

2. Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

**Вариант 2-09**

**Normal Bayes classifier**

**Задания**

1. Оценить ковариационные матрицы обучающих данных для каждого класса при различных предположениях: а) ковариационные матрицы всех классов равные, скалярные; б) равные, диагональные; в) различные, скалярные; г) различные, диагональные; д) равные; е) различные. Визуализировать матрицы для каждого случая.

2. Рассчитать классификационные очки обученных байесовских классификаторов (при предположениях а)–е) п.1) для примеров обучающей и тестовой выборок.

3. В исходном пространстве признаков изобразить области классов (закрасить разными цветами), формируемые каждым из обученных классификаторов. Нанести на диаграммы границы классов и данные из обучающей и тестовой выборок. На отдельной диаграмме изобразить все границы классов, формируемые построенными классификаторами.

4. Рассчитать точности (accuracy) построенных классификаторов на обучающей и тестовой выборках, представить данные в графическом виде (bar charts).

5. Для каждого классификатора построить micro-averaged и macro-averaged ROC-кривые и PR-кривые на обучающей и тестовой выборках и рассчитать micro-averaged и macro-averaged ROC AUC и PR AUC на обучающей и тестовой выборках.

6. Сделать выводы о влиянии предположений о модели обучающих данных на точность байесовской классификации.

**Указания**

Кросс-валидация: Holdout (70/30).

Априорные вероятности классов считать равными.

Классификационные очки классификаторов полагать равными апостериорным вероятностям классов.

**Литература**

1. Figueiredo, M. A. (2004). Lecture notes on bayesian estimation and classification. *Instituto de Telecomunicacoes-Instituto Superior Tecnico*, *60*.

2. Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

**Вариант 2-10**

**Regularized normal Bayes classifier**

**Задания**

1. Оценить ковариационные матрицы обучающих данных для каждого класса при различных предположениях: а) ковариационные матрицы всех классов равные, скалярные; б) равные; в) различные. Визуализировать матрицы для каждого случая.

2. Рассчитать классификационные очки регуляризованных байесовских классификаторов при различных параметрах регуляризации α и β. Визуализировать матрицы точности (accuracy) классификаторов на обучающей и тестовой выборках. По столбцам матриц – значения α, по строкам – значения β. Определить, при каких α и β классификатор обладает наибольшей обобщающей способностью.

3. В исходном пространстве признаков изобразить области классов (закрасить разными цветами), формируемые классификатором: а) обладающим наибольшей (по α, β) точностью на обучающей выборке; б) обладающим наибольшей (по α, β) точностью на тестовой выборке; в) при α = 1, β = 0; г) при α = 0, β = 1; в) при α = 0, β = 0. Нанести на диаграммы границы классов и данные из обучающей и тестовой выборок.

4. Рассчитать значения macro-averaged ROC AUC на обучающей и тестовой выборках для регуляризованных байесовских классификаторов при различных параметрах регуляризации α и β. Визуализировать матрицы значений macro-averaged ROC AUC на обучающей и тестовой выборках. По столбцам матриц – значения α, по строкам – значения β. Сравнить с матрицами точности (п.2).

5. В исходном пространстве признаков изобразить области классов (закрасить разными цветами), формируемые классификатором, обладающим наибольшим (по α, β) значением macro-averaged ROC AUC на обучающей и тестовой выборках. Нанести на диаграммы границы классов и данные из обучающей и тестовой выборок. Сравнить с границами, полученными в п.3.

6. Сделать выводы по влиянию параметров регуляризации на точность классификации и значение ROC AUC на обучающей и тестовой выборках.

**Указания**

Кросс-валидация: Holdout (70/30).

Априорные вероятности классов считать равными.

Классификационные очки классификаторов полагать равными апостериорным вероятностям классов.

В пп. 2,4 при визуализации матриц использовать heatmap либо поверхность (surface plots) и контурные диаграммы (contour plots).

**Литература**

1. Friedman, J. H. (1989). Regularized discriminant analysis. Journal of the American statistical association, 84(405), 165-175.

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

**Вариант 2-11**

**Bayesian document categorization: Bernoulli model**

**Задания**

1. Для каждой категории рассчитать среднее значение и с.к.о. (по документам) числа содержащихся в них различных (повторы не учитывать): а) слов; б) стоп-слов. Построить гистограммы распределения документов по числу содержащихся в них: а) различных слов; б) различных стоп-слов.

2. Сформировать словарь (из всех слов, встречающихся во всех документах), исключить из словаря стоп-слова.

3. Визуализировать бинарную матрицу вхождения слов в документы (по строкам – документы, по столбцам – слова из словаря, 1 – слово встречается в документе, 0 – не встречается). Привести 10 слов, наиболее часто встречающихся в документах из каждой категории. Для каждой категории представить графики зависимости относительного числа документов, содержащих заданное слово, от номера слова (слова отсортированы в порядке уменьшения числа содержащих его документов).

4. Рассчитать классификационные очки байесовского классификатора для каждого документа из обучающей и тестовой выборок. Построить матрицу ошибок (confusion matrix) классификатора. Для выделенных в п.3 наиболее часто встречающихся слов визуализировать матрицу рассчитанных классификационных очков (по строкам – категории, по столбцам – слова).

5. Рассчитать micro-averaged и macro-averaged показатели, основанные на таблице сопряженности: error rate, accuracy, sensitivity, specificity, precision, fall-out, F1-score, Cohen’s kappa.

6. Сделать выводы о качестве построенного байесовского классификатора документов.

**Указания**

Кросс-валидация: Holdout (70/30).

Классификационные очки классификаторов полагать равными апостериорным вероятностям классов.

Шапки документов (email, тема, дата, кол-во строк и пр.) не обрабатывать.

Знаки пунктуации, числа и специальные символы исключать.

**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.

**Вариант 2-12**

**Bayesian document categorization: Multinomial model**

**Задания**

1. Для каждой категории рассчитать среднее значение и с.к.о. (по документам) числа содержащихся в них: а) слов; б) стоп-слов (включая повторы). Построить гистограммы распределения документов по числу содержащихся в них: а)  слов; б) стоп-слов (включая повторы).

2. Сформировать словарь (из всех слов, встречающихся во всех документах), исключить из словаря стоп-слова.

3. Визуализировать матрицу частот вхождения слов в документы (по строкам – документы, по столбцам – слова из словаря). Привести 10 слов, наиболее часто встречающихся (суммарно) в документах из каждой категории. Для каждой категории представить графики зависимости относительной суммарной (по документам) частоты встречаемости слов от номера слова (слова отсортированы в порядке уменьшения частоты встречаемости).

4. Рассчитать классификационные очки байесовского классификатора для каждого документа из обучающей и тестовой выборок. Построить матрицу ошибок (confusion matrix) классификатора. Для выделенных в п.3 наиболее часто встречающихся слов визуализировать матрицу рассчитанных классификационных очков (по строкам – категории, по столбцам – слова).

5. Рассчитать micro-averaged и macro-averaged показатели, основанные на таблице сопряженности: error rate, accuracy, sensitivity, specificity, precision, fall-out, F1-score, Cohen’s kappa.

6. Сделать выводы о качестве построенного байесовского классификатора документов.

**Указания**

Кросс-валидация: Holdout (70/30).

Классификационные очки классификаторов полагать равными апостериорным вероятностям классов.

Шапки документов (email, тема, дата, кол-во строк и пр.) не обрабатывать.

Знаки пунктуации, числа и специальные символы исключать.

**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.

**Вариант 2-13**

**Non-parametric naïve Bayes classifier**

**Задания**

1. В предположении о независимости признаков построить графики восстановленных одномерных плотностей распределения каждого признака для каждого класса с использованием: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Ширину окон определить по правилу Сильвермана.

2. Рассчитать среднее значение и с.к.о. (по фолдам) точности (accuracy) обученного непараметрического байесовского классификатора на обучающей и тестовой выборках для случаев а)–г) п.1.

3. Построить графики зависимости среднего значения и с.к.о. (по фолдам) точности (accuracy) обученного байесовского классификатора на обучающей и тестовой выборках от коэффициента пропорциональности λ (отношение ширины парзеновского окна к ширине Сильвермана) для случаев а)–г) п.1. Для каждого типа окна определить его вектор ширин, при котором байесовский классификатор обладает наибольшей обобщающей способностью.

4. Обучить байесовский классификатор в предположении о нормальности распределения данных всех классов с равными диагональными ковариационными матрицами. Рассчитать среднее значение и с.к.о. (по фолдам) точности (accuracy) классификатора и сравнить со значениями, полученными в п.3.

5. Построить ROC-кривые и PR-кривые и рассчитать ROC AUC и PR AUC на обучающей и тестовой выборках для непараметрического классификатора, обладающего наибольшей точностью на тестовой выборке (по результатам п.3). Сравнить с ROC-кривыми и PR-кривыми параметрического классификатора (п.4).

6. Сделать выводы о влиянии типа окна и ширины окна на качество непараметрической байесовской классификации.

**Указания**

Кросс-валидация: 5-fold.

Априорные вероятности классов считать равными.

Для визуализации исходных данных использовать диаграммы рассеяния в подпространствах, формируемых каждой парой признаков.

В п.3 вектор ширин парзеновских окон выбирать пропорциональным вектору ширины Сильвермана: .

**Литература**

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

2. Pérez, A., Larrañaga, P., & Inza, I. (2009). Bayesian classifiers based on kernel density estimation: Flexible classifiers. *International Journal of Approximate Reasoning*, *50*(2), 341-362.

**Вариант 2-14**

**Non-parametric Bayes classifier**

**Задания**

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

2. Рассчитать среднее значение и с.к.о. (по фолдам) точности (accuracy) обученного непараметрического байесовского классификатора на обучающей и тестовой выборках для случаев а)–г) п.1.

3. Визуализировать зависимость среднего значения и с.к.о. (по фолдам) точности (accuracy) обученного байесовского классификатора на обучающей и тестовой выборках от коэффициента пропорциональности λ (отношение ширины парзеновского окна к ширине Сильвермана) для случаев а)–г) п.1. Для каждого типа окна определить вектор ширин, при котором байесовский классификатор обладает наибольшей обобщающей способностью.

4. Обучить байесовский классификатор в предположении о нормальности распределения данных всех классов. Рассчитать среднее значение и с.к.о. (по фолдам) точности (accuracy) классификатора и сравнить с точностями, полученными в п.3.

5. В исходном пространстве признаков изобразить области классов (закрасить разными цветами), формируемые: а) непараметрическим байесовским классификатором, обладающим наибольшей точностью на обучающей и тестовой выборках; б) нормальным байесовским классификатором, обученным в п.4. Нанести на диаграммы границы классов и данные из обучающей и тестовой выборок. Сравнить формируемые границы классов.

6. Сделать выводы о влиянии типа окна и ширины окна на качество непараметрической байесовской классификации.

**Указания**

Кросс-валидация: 5-fold.

Априорные вероятности классов считать равными.

В п.3 вектор ширин парзеновских окон выбирать пропорциональным вектору ширины Сильвермана: .

**Литература**

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

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

3. Pérez, A., Larrañaga, P., & Inza, I. (2009). Bayesian classifiers based on kernel density estimation: Flexible classifiers. *International Journal of Approximate Reasoning*, *50*(2), 341-362.

**Вариант 2-15**

**Normal vs non-parametric Bayes classifier**

**Задания**

1. Оценить ковариационные матрицы обучающих данных для каждого класса при различных предположениях: а) ковариационные матрицы всех классов равные, скалярные; б) равные, диагональные; в) различные, скалярные; г) различные, диагональные; д) равные; е) различные. Визуализировать матрицы для каждого случая.

2. Рассчитать классификационные очки обученных байесовских классификаторов (при предположениях п.1) для примеров обучающей и тестовой выборок.

3. Рассчитать точности (accuracy) построенных классификаторов на обучающей и тестовой выборках, представить данные в графическом виде (bar charts).

4. В предположении о независимости парзеновских ядер визуализировать восстановленные двумерные плотности распределения для каждого класса. Использовать гауссово окно, ширину окна определить по правилу Сильвермана.

5. Построить графики зависимости среднего значения и с.к.о. (по фолдам) точности (accuracy) обученного байесовского классификатора на обучающей и тестовой выборках от коэффициента пропорциональности λ (отношение ширины парзеновского окна к ширине Сильвермана). Определить ширину окна, при которой байесовский классификатор обладает наибольшей точностью на обучающей и тестовой выборках.

6. Сделать выводы о точности в параметрической и непараметричской байесовской классификации.

**Указания**

Кросс-валидация: 5-fold.

Априорные вероятности классов считать равными.

Классификационные очки классификаторов полагать равными апостериорным вероятностям классов.

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

**Литература**

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

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

**Вариант 2-16**

**ECOC design for non-parametric Bayes classifier**

**Задания**

1. В предположении о независимости парзеновских ядер визуализировать восстановленные двумерные плотности распределения для каждого класса. Использовать гауссовы окна, ширину окон определить по правилу Сильвермана.

2. Обучить непараметрические байесовские бинарные классификаторы при различных способах ECOC-кодирования: а) OVA; б) OVO; в) полное бинарное кодирование; г) полное тернарное кодирование.

3. Для каждого из полученных в п.1 наборов базовых классификаторов рассчитать классификационные очки многоклассового классификатора, используя различные способы ECOC-декодирования: а) взвешенное; б) невзвешенное.

4. В исходном пространстве признаков изобразить области классов (закрасить разными цветами), формируемые каждым из полученных в п.2 многоклассовым классификатором (всего 8 классификаторов). Нанести на диаграммы границы классов и данные из обучающей выборки. На отдельной диаграмме изобразить все границы классов, формируемые построенными многоклассовыми классификаторами.

5. Построить графики зависимости показателей качества (accuracy, micro-averaged и macro-averaged ROC AUC и PR AUC) построенных многоклассовых классификаторов на обучающей и тестовой выборках от коэффициента пропорциональности λ (отношение ширины парзеновского окна к ширине Сильвермана). Определить для каждого классификатора ширину окна, при которой наблюдается наилучшая обобщающая способность.

6. Сделать вывод о влиянии способов ECOC-кодирования, ECOC-декодирования и ширины парзеновского окна на точность многоклассовой классификации.

**Указания**

Кросс-валидация: Holdout (70/30).

Априорные вероятности классов для каждого бинарного классификатора считать равными.

В п.2. для оценки плотности распределения данных из нескольких классов использовать модель смеси распределений: плотность распределения данных из 1, …, *k* классов полагать равной , где  – оцененная функция плотности распределения *i*-го класса, *ni* – объём выборки *i*-го класса.

Классификационные очки бинарных классификаторов считать по формуле , где *pk* – апостериорная вероятность *k*-го класса.

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

**Литература**

1. Escalera, S., Pujol, O., & Radeva, P. (2009). Separability of ternary codes for sparse designs of error-correcting output codes. *Pattern Recognition Letters*, *30*(3), 285-297.

**Вариант 2-17**

**ECOC random coding for non-parametric Bayes classifier**

**Задания**

1. В предположении о независимости признаков построить графики восстановленных одномерных плотностей распределения каждого признака для каждого класса с использованием: а) прямоугольного окна; б) гауссова окна; в) окна Епанечникова; г) треугольного окна. Ширину окна определить по правилу Сильвермана.

2. Обучить непараметрические байесовские бинарные классификаторы при случайном ECOC-кодировании: а) бинарном (dense); б) тернарном (sparse). Число бинарных классификаторов, участвующих в кодировании, задать произвольно.

3. Для каждого из полученных в п.1 наборов базовых классификаторов рассчитать классификационные очки многоклассового классификатора, используя различные способы ECOC-декодирования: а) взвешенное; б) невзвешенное.

4. Визуализировать матрицу точностей (accuracy) обученных в п.2 многоклассовых классификаторов на обучающей и тестовой выборках: по строкам матрицы – число бинарных классификаторов, участвовавших в случайном ECOC-кодировании, по строкам – коэффициент пропорциональности λ (отношение ширины окна к ширине Сильвермана). Определить число бинарных классификаторов и вектор ширин окон, при которых наблюдается наилучшая точность классификации на тестовой выборке.

5. В исходном пространстве признаков изобразить области классов (закрасить разными цветами), формируемые многоклассовым классификатором, обладающим наибольшей точностью (accuracy) на обучающей выборке, при а) взвешенном; б) невзвешенном ECOC-декодировании. Нанести на диаграммы границы классов и данные из обучающей выборки.

6. Сделать вывод о влиянии числа бинарных классификаторов, участвующих в случайном ECOC-кодировании, и ширины парзеновского окна на точность классификации.

**Указания**

Кросс-валидация: Holdout (70/30).

Априорные вероятности классов для каждого бинарного классификатора считать равными.

В п.2. для оценки плотности распределения данных из нескольких классов использовать модель смеси распределений: плотность распределения данных из 1, …, *k* классов полагать равной , где  – оцененная функция плотности распределения *i*-го класса.

Классификационные очки бинарных классификаторов считать по формуле , где *pk* – апостериорная вероятность *k*-го класса.

Столбцы матрицы случайного бинарного (тернарного) кодирования отбирать случайным образом (без повторений) из столбцов матрицы полного бинарного (тернарного) кодирования,

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

**Литература**

1. Escalera, S., Pujol, O., & Radeva, P. (2009). Separability of ternary codes for sparse designs of error-correcting output codes. *Pattern Recognition Letters*, *30*(3), 285-297.