Задание по теме:

Непараметрический байесовский классификатор

1. Найти размеченную выборку данных для задачи многоклассовой классификации либо сгенерировать свою выборку (например, из распределения $C_k Z_k / ||Z_k|| + 0.1 Z_k$, где Z_k имеет двумерное круговое нормальное распределение, C_k – константа, своя для каждого класса).
2. Провести разведочный анализ данных (построить гистограммы распределения признаков в каждом классе, диаграммы рассеяния и box-and-whisker, оценить статистические характеристики выборки, соответствие распределения данных нормальному распределению и т.д.).
3. При необходимости провести предобработку реальных данных (устранить дубликаты и выбросы, восстановить пропущенные значения, удалить неинформативные признаки и т.д.).
4. Обучить непараметрические байесовские классификаторы.
   4.1. 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.
   4.2. Рассчитать среднее значение и с.к.о. (по фолдам кросс-валидации) точности (accuracy) обученного непараметрического байесовского классификатора на обучающей и тестовой выборках для случаев а)–г).
   4.3. 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). Для каждого типа окна и каждого признака определить ширину окна, при которой байесовский классификатор обладает наибольшей обобщающей способностью.
   4.4. В исходном пространстве признаков (либо в нескольких проекциях) 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.
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. 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 с аналогичными значениями для непараметрического байесовского классификатора.

7. Провести исследования построенных моделей: оценить влияние априорных вероятностей классов на границы и показатели качества классификации, сравнить границы классов параметрического и непараметрического классификаторов и пр.
8. Сделать выводы о влиянии ширины и вида парзеновского окна на точность непараметрической байесовской классификации.
9. Обучить модель логистической регрессии и сравнить показатели точности непараметрического байесовского классификатора и логистической регрессии.
10. Оформить отчет о результатах проведенных исследований.