домашнее задание: нормальный дискриминантный анализ

5 марта 2013 г.

крайний срок сдачи - 11.03.2013, 23.59

1) Два класса имеют n-мерные нормальные распределения со средними μ_1 и μ_2 и одинаковыми ковариационнмыи матрицами Σ. Априорные вероятности и значимости классов равны: $P_1 = P_2, \lambda_1 = \lambda_2$. Докажите, что:
a) линии уровня плотности распределения каждого из классов — эллипсы. Многомерный эллипс — такое множество точек x, которое при переходе в новую систему координат описыается уравнением $\sum_{i=1}^{n} \frac{x_i^2}{a_i^2} = 1$. Что можно сказать о величинах a_i?
b) разделяющая поверхность пройдет через середину отрезка, соединяющего центры
классов $(\mu_1 + \mu_2)/2$.
c) разделяющая поверхность в точке $(\mu_1 + \mu_2)/2$ касается уровней плотности распределений обоих классов.

2) Два класса имеют двумерные нормальные распределения. Может ли граница оптимального байесовского классификатора представлять собой эллипс? Если да, приведите пример (т.е. задайте $P_1, P_2, \lambda_1, \lambda_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2$), если нет — докажите.

3) Мини-практическое задание.
На практике, при решении задачи классификации, возникает проблема мультиколлинеарности признаков. Оценка матрицы ковариации $\hat{\Sigma}$ получается плохообусловленной, из-за чего подстановочный алгоритм обладает низкой обобщающей способностью. Один из методов борьбы с этой проблемой — регуляризация матрицы $\hat{\Sigma}$. Вместо исходной оценки $\hat{\Sigma}$ рассматривают $\hat{\Sigma} + \tau \times I$, где I — единичная матрица.

Вам предоставляются данные для обучения train.csv и данные для прогноза test.csv. Каждая строка содержит признаки одного объекта, разделенные запятой ','. В train.csv столбцов на один больше, так как последний столбец в train.csv соответсвует классу объекта (0 или 1). Классы объектов из test.csv не известны и подлежат прогнозу. Запрограммируйте подстановочный алгоритм, сделав предположение о нормальности распределения объектов в обоих классах Данные подобраны таким образом, что матрица $\hat{\Sigma}$ получится плохообусловленной, поэтому ее необходимо регуляризировать описанным выше методом. Предложите способ выбора τ. Семинаристу вы должны предоставить:

- код программы
- отчет, в котором отразить способ выбора параметра τ.
- файл с прогнозом тестовых данных *test.csv*. В файле должно быть ровно 2000 строк, на i-й строке которого должен быть записан либо 0 либо 1.