Задание 2 (I курс, весенний семестр 2010г.)

Задача 1 (обязательная для всех).
Задачу приближения функции регрессии $M(y\,|\,x)$ в классе функций $F=\{f(x,a)\}$, где a - вектор параметров, можно понимать как задачу минимизации на F функционала среднего риска $J_m(a)=\int (y-f(x,a))^2 dP(x,y)$. Функционал эмпирического риска при этом имеет вид $J_e(a)=\frac{1}{l}\sum_{i=1}^{l}(y_i-f(x_i,a))^2$, где $(x_1,y_1,...,x_l,y_l)$ - независимая выборка значений пары (x,y), имеющей распределение $P(x,y)$. При построении равномерной по классу F оценки величины уклонения эмпирического риска от среднего риска рассматривают множества вида $A_{a,c}=\{x,y:(y-f(x,a))^2<c\}$.

Оценить функцию роста системы $A_{a,c}$ (параметры c и a пробегают все возможные значения) в задаче построения регрессии методом минимизации эмпирического риска в классе

- полиномов степени не выше n (x и y - скаляры)
- линейных функций вида $f(x,a)=\sum_{i=1}^{n}a_i z_i$, где z_i - i-ая координата вектора x.

Задача 2.
Каждый объект, принадлежащий одному из 3 классов, обозначим их S, B, U, описывается 3 характеристиками b, w, h, имеющими в каждом из классов нормальное распределение.

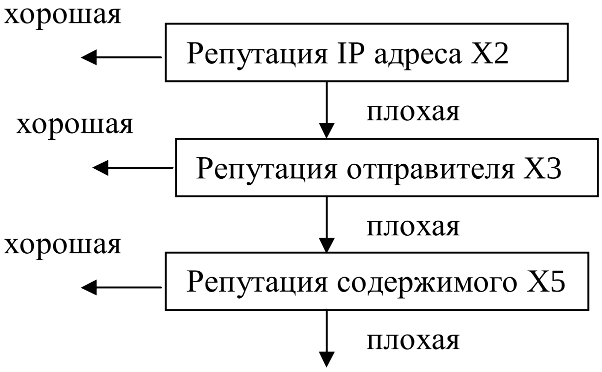
1) Записать следующее правило классификации: объект относится к тому классу, апостериорная вероятность которого максимальна.
2) Построить решающее правило для распознавания классов S, B, U при единичных ковариационных матрицах в каждом из классов и средних значениях в классах равных $m_S=\frac{1}{15}(90\quad 90\quad 60)$, $m_B=\frac{1}{15}(90\quad 60\quad 90)$, $m_U=\frac{1}{15}(60\quad 90\quad 90)$.
   Априорные вероятности классов считать свободными параметрами.
3) Дать геометрическую интерпретацию полученному правилу классификации.

Задача 3.
Репутационный метод выявления спама основан на признаках

- X1 Репутация домена
- X2 Репутация адреса IP
- X3 Репутация отправителя
- X4 Репутация получателя
- X5 Репутация содержимого письма

Будем считать перечисленные признаки двоичными (0-плохая репутация; 1 – хорошая репутация) и обозначим $P(S\,|\,Xi=0)$ вероятность того, что сообщение, для которого признак Xi равен нулю, является спамом, а через $P(S\,|\,Xi=1)$ вероятность того, что сообщение, для которого признак Xi равен единице, является спамом.
Допустим, что условная вероятность того, что сообщение с набором признаков X1, X2, X3, X4, X5 является спамом, задаётся соотношением
$P(S\,|\,X1,X2,X3,X4,X5)=P(S\,|\,X1)P(S\,|\,X2)P(S\,|\,X3)P(S\,|\,X4)P(S\,|\,X5)$

- записать выражение для вероятности того, что пришедшее письмо является спамом
- записать выражение для вероятности того, что пришедшее письмо является спамом, если в сети равновероятно присутствуют адреса IP с плохой и с хорошей репутацией. Все отправители с плохой репутацией принадлежат только адресам IP с плохой репутацией, но среди отправителей, принадлежащих адресам IP с плохой репутацией, равновероятно присутствуют и отправители с хорошей репутацией. Все письма с плохой репутацией отправляются только отправителями с плохой репутацией, но содержания писем, полученных от отправителя с плохой репутацией, равновероятно могут иметь как плохую, так и хорошую репутацию (см. рисунок).



Задача 4.

Пусть случайные векторы X и Y связаны соотношением $Y = AX + e$, где A – фиксированная матрица, X имеет нормальное распределение с математическим ожиданием X_0 и ковариационной матрицей K_x, а e имеет нормальное распределение с нулевым математическим ожиданием и ковариационной матрицей K_e. Записать

1. выражение для апостериорного распределения вектора X при заданном векторе Y
2. выражение для апостериорного среднего значения вектора X при заданном векторе Y.