Задача 1.

В байесовском подходе мы относим наш вектор x к классу $a(x)$ по правилу:

$$a(x) = argmax_y P(y|x).$$

По формуле Байеса и из предположения, что все признаки независимы, мы получаем формулу:

$$a(x) = argmax_y \frac{P(y)\prod_{k=1}^{n} P(x^{(k)}|y)}{P(x)}.$$

При этом можно не рассматривать знаменатель, так как максимизация идет по x, а также выкинуть $P(y)$, так как по условию $P(y)$ равны для любого y.

Поэтому рассмотрим

$$P(x|y) = \prod_{k=1}^{n} P(x^{(k)}|y) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\sum_{k=1}^{n}(x^{(k)}-\mu_{yk})^2}{2\sigma^2}}.$$

Отсюда следует, что для максимизации $P(x|y)$ мы должны минизировать $\sum_{k=1}^{n}(x^{(k)}-\mu_{yk})^2$. А эта величина как раз и есть расстояние до μ_y.

Задача 2.

Мы рассматриваем "треугольный ROC_AUC". ROC-кривая при этом строится по трем точкам: (0, 0), (1, 1) и (FPR, TPR), где точка (FPR, TPR) строится данным классификатором.

Площадь под ROC-кривой можно найти как сумму площадей прямоугольного треугольника и трапеции:

$$S = \frac{1}{2} FPR \cdot TPR + (1 - FPR)\frac{1 + TPR}{2} = \frac{TPF + 1 - FPR}{2}.$$

Таким образом, чтобы найти $E(S)$, нам необходимо знать E(TPR) и E(FPR). Обозначим через N_1 — число объектов класса "1", а N_0 — число объектов класса "0".

По определению:

$$TPR = \frac{TP}{TP + FN},\ FPR = \frac{FP}{FP + TN}.$$

Тогда

$$E(TPR) = E\left(\frac{TP}{TP + FN}\right) = \frac{E(TP)}{N_1} = \frac{E\left(\sum\limits_{x:y_x=1} I(a(x) = 1)\right)}{N_1} = \frac{N_1 p}{N_1} = p.$$

$$E(FPR) = E\left(\frac{FP}{FP+TN}\right) = \frac{E(FP)}{N_0} = \frac{E\left(\sum\limits_{x:y_x=1} I(a(x)=0)\right)}{N_0} = \frac{N_0 p}{N_0} = p.$$

Теперь мы можем найти среднее значение площади:

$$E(S) = E\left(\frac{TPF + 1 - FPR}{2}\right) = \frac{p+1-p}{2} = \frac{1}{2},$$

что нам и требовалось.

Задача 3.

Обозначим через y — настоящий класс x, а через y_n — класс ближайшего соседа x_n. Так как мы рассматриваем случай бинарной классификации, то уравнение для E_n будет иметь вид:

$$E_N = P(y \neq y_n) = P(x=1, y_n=0) + P(x=0, y_n=1) = P(1|x)P(0|x_n) + P(0|x)P(1|x_n).$$

Осуществим в этой формуле предельный переход по n, тогда

$$E_N \to 2P(1|x)P(0|x).$$

Так как $P(1|x) + P(0|x) = 1$, то $P(1|x) = 1 - P(0|x)$. Также учитывая, что $E_B = min\{P(1|x), P(0|x)\}$ получаем

$$E_N \to 2P(1|x)P(0|x) = 2(1-P(0|x))P(0|x) = 2(1-P(1|x))P(1|x) = 2E_B(1-E_B) \leq 2E_B.$$

А это нам и требовалось доказать.