P2

(a)

$$
\begin{aligned}
2c \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}} &< \frac{2}{\sqrt{2\pi}} \cdot e^{-2(x-2)^2} \\
\ln c - \frac{x^2}{2} &< -2(x-2)^2 \\
3x^2 - 16x + 16 + 2\ln c &< 0 \\
\frac{1}{6}(8 - \sqrt{64 - 24\ln c}) < x &< \frac{1}{6}(8 + \sqrt{64 - 24\ln c})
\end{aligned}
$$

(b) None, because when $x > e^{\frac{8}{3}}$, previous function has no real root.

(c) seperate the data based on their labels
sort the data based on their values in the first columns
find out how many data with each label we need (each = n/10)
find out how many data should we put in one group (floor(number of data with such label / each))
find out the mean of each group
form the final prototype data set
form the prototype label set
test

(d)

Error rate	1000	2000	4000	8000
P1	11.34%	9.06%	7.34%	5.65%
P2	9.18%	8.29%	7.22%	6.18%

P3

(1) The probability that two balls have different color is

$$P = P(c_1)\cdot(1-P(c_1))+P(c_2)\cdot(1-P(c_2))+P(c_3)\cdot(1-P(c_3))+P(c_4)\cdot(1-P(c_4))+P(c_5)\cdot(1-P(c_5))$$

Because the two picks are independent, so the probability that another color is picked after certain color is picked is $P(c_i) \cdot (1 - P(c_i))$.

And the probability that two picks are different is the sum of the probability that each color is picked first.

(2) We know $c_1 + c_2 + c_3 + c_4 + c_5 = 100$

$$P = P(c_1)\cdot(1-P(c_1))+P(c_2)\cdot(1-P(c_2))+P(c_3)\cdot(1-P(c_3))+P(c_4)\cdot(1-P(c_4))+P(c_5)\cdot(1-P(c_5))$$

$$P = \frac{c_1}{100} \cdot (1 - \frac{c_1}{100}) + \ldots + \frac{c_5}{100} \cdot (1 - \frac{c_5}{100})$$

$$P = 1 - \left((\frac{{c_1}^2}{10000}) + \ldots + (\frac{{c_5}^2}{10000}) \right)$$

By the CauchySchwarz inequality,

$${c_1}^2 + {c_2}^2 + {c_3}^2 + {c_4}^2 + {c_5}^2 \geq \frac{(c_1 + c_2 + c_3 + c_4 + c_5)^2}{5} = \frac{100^2}{5} = 2000$$

$${c_1}^2 + {c_2}^2 + {c_3}^2 + {c_4}^2 + {c_5}^2 = \frac{(c_1 + c_2 + c_3 + c_4 + c_5)^2}{5} \text{ when } c_1 = c_2 = c_3 = c_4 = c_5 = \frac{100}{5} = 20$$

$$P = 1 - 0.2$$

$$P = 0.8$$

There should be 20 balls with each of the five colors.