Task 1

$$P_\theta = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$L(S;\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - m\log\left(\sigma\sqrt{2\pi}\right)$$

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \hat{\sigma} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(x_i-\hat{\mu}\right)^2}$$

Покажем, что $\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{m-1}{m}\sigma^2$.

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x_i^2 - 2x_i\hat{\mu} + \hat{\mu}^2\right)\right] = \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x_i^2 - \hat{\mu}^2\right] = \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[x_i^2\right] - \mathbb{E}\left[\hat{\mu}^2\right] = \\
&= \mathbb{E}\left[x^2\right] - \mathbb{E}\left[\hat{\mu}^2\right] = \\
&= \left(\mu^2+\sigma^2\right) - \left(\mu^2 + \mathbb{D}[\hat{\mu}]\right) = \\
&= \sigma^2 - \frac{1}{m^2}\mathbb{D}\left[\sum_{i=1}^{m} x_i\right] = \sigma^2 - \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{D}\left[x_i\right] = \sigma^2 - \frac{1}{m}\sigma^2 = \\
&= \frac{m-1}{m}\sigma^2
\end{aligned}$$

Таким образом, оценка максимального правдоподобия для нормально распределенной случайной является смещенной.

Task 2

Добавим в исходную выборку один элемент класса 1 и один элемент класса 0. Обозначим их x_{m+1}, x_{m+2} соответственно. Отметим, что речь идет о распределении Бернулли, при этом $P\left[X=1\right] = \theta$. Учитывая, что

$$L\left(S;\theta\right) = \log\left(\prod_{i=1}^{m} P_\theta(x_i)\right) = \sum_{i=1}^{m}\log\left(P_\theta(x_i)\right),\ \hat{\theta} \in \underset{\theta}{argmax}\ L\left(S;\theta\right);$$

$$l(\theta;x) = -\log\left(P_\theta\left[x\right]\right);$$

$$\underset{\theta}{argmin}\sum_{i=1}^{m}\left(-\log\left(P_\theta\left[x_i\right]\right)\right) = \underset{\theta}{argmax}\sum_{i=1}^{m}\left(\log\left(P_\theta\left[x_i\right]\right)\right);$$

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{m+2}\sum_{i=1}^{m+2} x_i = \frac{1}{m+2}\left(1 + \sum_{i=1}^{m} x_i\right);$$

получаем

$$L\left(S;\theta\right) = \sum_{i=1}^{m+2}\log\left(P_\theta(x_i)\right) = \sum_{i=1}^{m}\log\left(P_\theta(x_i)\right) + \log\left(\theta\right) + \log\left(1-\theta\right).$$

Т.е. минимизация регуляризированной целевой функции для исходной выборки эквивалентна минимизации эмперического риска для расширенной выборки.

Получим верхнюю оценку для $|\hat{\theta} - \theta^\star|$.

$$|\hat{\theta}-\theta^\star| \leq |\hat{\theta}-\mathbb{E}[\hat{\theta}]| + |\mathbb{E}[\hat{\theta}]-\theta^\star|,\ \mathbb{E}[\hat{\theta}] = \frac{1+m\theta^\star}{m+2}.$$

Тогда

$$|\hat{\theta}-\mathbb{E}[\hat{\theta}]| = \frac{m}{m+2}\left|\frac{1}{m}\sum_{i=1}^{m} x_i - \theta^\star\right|,\ |\mathbb{E}[\hat{\theta}]-\theta^\star| = \left|\frac{1-2\theta^\star}{m+2}\right| \leq \frac{1}{m+2}$$

Используя неравенство Хефдинга, получаем

$$\mathbb{P}\left[|\hat{\theta}-\theta^\star| \geq \frac{1}{m+2}+\epsilon\right] = \mathbb{P}\left[|\hat{\theta}-\theta^\star| \geq \varepsilon\right] \leq 2e^{-m\varepsilon^2}.$$

Тогда с вероятностью $1-\delta$

$$|\hat{\theta}-\theta^\star| \leq \sqrt{\frac{\log\frac{1}{\delta}}{m}} = O\left(\frac{1}{\sqrt{m}}\right).$$

Task 3

Шаг M метода *soft k-means* заключается в максимизации ожидаемого логарифма правдоподобия.

$$\{c \in \mathbb{R}^k : \sum_{i=1}^{k} c_i = 1,\ c_i \geq 0\}.$$

$$\underset{\theta}{max} \sum_{i=1}^{m}\sum_{y=1}^{k} P_{\theta^{(t)}}\left[Y=y|X=x_i\right]\left(\log(c_y) - \frac{1}{2}\parallel x_i-\mu_y \parallel^2\right).$$

$$\mu_y = \frac{\sum_{i=1}^{m} P_{\theta^{(t)}}[Y=y|X=x_i]x_i}{\sum_{i=1}^{m} P_{\theta^{(t)}}[Y=y|X=x_i]}.$$

Обозначим $P_{\theta^{(t)}}$ как P, а также $\nu_y = \sum_{i=1}^{m} P\left[Y=y|X=x_i\right]$, тогда получаем, что исходная максимизация эквивалентна следующей задаче оптимизации:

$$\max_{c\in\mathbb{R}^k} \sum_{y=k}^{k} \nu_y \log(c_y),\ c_y \geq 0, \sum_y c_y = 1.$$

Пусть $c^\star = \frac{\nu}{\sum_y \nu_y}$. Учитывая, что $\nu_y \geq 0$ и $\sum_y c^\star_y = 1$, тогда $c^\star$ - вектор вероятности.

$$-D_{RE}(c^\star \parallel c) = \sum_y c^\star_y \log(\frac{c_y}{c^\star_y}) = \sum_y c^\star_y\left(\log(c_y) - \log(c^\star_y)\right) =$$
$$= Z_1\left(\sum_y \nu_y \log(c_y) + Z_2\right),$$

т.е. минимизация $D_{RE}(c^\star \parallel c)$ при условии $c_y \geq 0, \sum_y c_y = 1$ эквивалентна исходной задаче.

Так как $D_{RE}[P \parallel P_\theta] = \sum_x P[x]\log\left(\frac{P[x]}{P_\theta[x]}\right)$ и true risk минимален, а $L(x,\theta)$ достигает максимума при $P_\theta = P$.

Найдем данный оптимальный параметр $\theta^\star$.

$$L(x,\theta) = \sum_{i=1}^{m}\sum_{y=1}^{k} P_{\theta^{(t)}}\left[Y=y|X=x_i\right]\left(\log(c_y) - \frac{1}{2} \parallel x_i - \mu_y \parallel^2\right)$$
$$\frac{\partial L}{\partial \mu_y} = \sum_{i=1}^{m} P_{\theta^{(t)}}\left[Y=y|X=x_i\right] \parallel x_i - \mu_y \parallel = 0$$
$$\mu_y = \frac{\sum\limits_{i=1}^{m} P_{\theta^{(t)}}[Y=y|X=x_i]x_i}{\sum\limits_{i=1}^{m} P_{\theta^{(t)}}[Y=y|X=x_i]}.$$

Воспользуемся правилом множителей Лагранжа

$$L(x,\theta) + \lambda\left(\sum_{y=1}^{k} c_y - 1\right)$$
$$\sum_{y=1}^{k}\left(\frac{\partial L}{\partial c_y} + \lambda\right) = 0 \Rightarrow \lambda = \sum_{y'=1}^{k}\sum_{i=1}^{m} P_{\theta^{(t)}}\left[Y=y'|X=x_i\right]$$
$$c_y = \frac{\sum\limits_{i=1}^{m} P_{\theta^{(t)}}[Y=y|X=x_i]}{\sum\limits_{y'=1}^{k}\sum\limits_{i=1}^{m} P_{\theta^{(t)}}[Y=y'|X=x_i]} \Rightarrow c^\star = \frac{\nu}{\sum\limits_{y}\nu_y}$$

Т.е. функция правдоподобия достигает максимума при $\theta^\star = \{\mu^\star, c^\star\}$.