Квиз

Пусть $X_1, \ldots, X_n$ – выборка независимых случайных величин, каждая из которых принадлежит к одному из двух кластеров. В k-ом кластере наблюдения распределены с функцией вероятности или функцией плотности $p_k(x|\theta_k)$, где θ_k – вектор неизвестных параметров. Пусть вероятность того, что наблюдение принадлежит первому кластеру, равна γ.

Обозначим за θ вектор, в который последовательно собраны неизвестные параметры для каждого из кластеров, а также γ:

$$\theta := \begin{pmatrix} -\theta_1- & -\theta_2- & \gamma \end{pmatrix}$$

Введите подходящие латентные переменные и выведите формулы для шагов EM-алгоритма (E-шаг – чему равно $p(Z|X, \theta_{old})$, M-шаг – формулы обновления $\theta_{new} = ...$), если

Задача 1

p_k – функция вероятности распределения Бернулли $\mathrm{Bern}(\alpha_k)$.

Распределение Бернулли: $P(X = 1) = \alpha, P(X = 0) = 1 - \alpha$

Задача 2

p_k – функция вероятности биномиального распределения $\mathrm{Bin}(3, \alpha_k)$.

Биномиальное распределение: $P(X = k) = C_n^k \cdot \alpha^k \cdot (1 - \alpha)^{n-k}$

Задача 3

p_k – функция плотности экспоненциального распределения $\exp(\lambda_k)$.

Экспоненциальное распределение: $f_X(x) = \lambda e^{-\lambda x}$

Задача 4

p_k – функция вероятности распределения Пуассона $\mathrm{Pois}(\lambda_k)$.

Распределение Пуассона: $P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$

Задача 5

p_k – функция вероятности геометрического распределения $\mathrm{Geom}(\alpha_k)$.

Геометрическое распределение: $P(X = k) = (1 - p)^k \cdot p$

Подготовка к контрольной работе

Теория информации

$$H(X) = -\sum_i p_i \log_2(p_i)$$

$$H(X) = -\int_a^b f(x)\log_2(f(x))dx$$

$$H(X,Y) = -\sum_{x,y} p(x,y)\log_2 p(x,y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$CE(X||Y) = -\sum p(x)\log_2 q(x)$$

$$H(X|Y) = H(X,Y) - H(Y)$$

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

$$D_{KL}(X\|Y) = CE(X\|Y) - H(X) = \sum_{x\in X} p(x)\log_2 \frac{p(x)}{p(y)}$$

Множественная проверка гипотез

	H_0 не отвергается	H_0 отвергается	Итого
H_0 верна	V	U	m_0
H_0 неверна	S	T	$m - m_0$
Итого	R	$m - R$	m

- FWER $= P\{U > 0\} = E[I\{U > 0\}] = \frac{m_0\cdot\alpha}{m}$
- FDR $= E[\frac{U}{max\{U+T,1\}}] =$
- поправка Бонферрони $\alpha_{ind} = \frac{\alpha}{m}$
- поправка Холма-Бонферрони $\alpha_k = \frac{\alpha}{m+1-k}$
- процедура Бенджамини-Хокберга $\alpha_k = \frac{\alpha\cdot k}{m}$

2.1 Энтропия нормального распределения (КР-2021, №6)

Условие задачи

Величина X имеет нормальное распределение $N(\mu,\sigma^2)$, а величина Y — другое распределение с математическим ожиданием μ и дисперсией σ^2.

Докажите, что $H(X) \geq H(Y)$ или приведите контр-пример.

Решение

По сути нам необходимо доказать, что у нормального распределения энтропия максимальна в классе распределений с одинаковой дисперсией. Доказательство этого факта приведено в английской википедии

2.2 Производная функции правдоподобия (КР-2021, №2)

Условие

Рассмотрим логарифм функции правдоподобия $l = 2\ln a + 4\ln b - 10a + 12b - 10$. Кажется, что если посчитать производные этой функции по a и b и взять математическое ожидание этих производных, то получается не ноль, хотя на лекции точно доказывалось, что $E(l_{\theta}^{'}) = 0$. Объясните это противоречие

Решение

l - это логарифм функции правдоподобия, то есть $l(x|\theta) = \ln L(x|\theta) = \ln \prod\limits_{i}^{n} P(x|\theta)$. Попробуем восстановить исходную функцию плотности для имеющегося логарифма правдоподобия:

$$L(\cdot) = \exp 2\ln a + 4\ln b - 10a + 12b - 10 = a^2 \cdot b^4 \cdot \exp(-10a + 12b - 10)$$

Видим, что получившаяся исходная функция правдоподобия никак не зависит от реализации выборки, то есть по сути является константой, в точке максимального правдоподобия ($\hat{a}_{ML} = 0.2, \hat{b}_{ML} = -\frac{1}{3}$) равной $0.2^2 \cdot 13^4 \cdot \exp(-2 + 4 - 10)$, что явлется предельно малым числом, никак не удовлевторяющим базовым свойствам вероятностной меры. Таким образом, можем сделать вывод, что максимизируемая функция правдоподобия не является функцией правдоподобия как таковой и поэтому можем сделать вывод, что доказанное на лекции никак не противоречит рассмотренному случаю.

2.3 Проверка гипотез для функции от параметра (КР-2021, №3)

Условие задачи

Пусть $X_1, ..., X_n$ – независимые случайные величины из распределения с функцией плотности или функцией вероятности $p(x|\theta)$. Обозначим как $I_{\theta}(\theta)$ информацию Фишера для задачи поиска $\hat{\theta}_{ML}$. Добрый волшебник Евгений решает ввести новый параметр μ, такой что $\theta = \psi(\mu)$, где ψ – дифференцируемая функция.

Обозначим как $I_{\mu}(\mu)$ информацию Фишера в терминах μ.

- Докажите, что $I_{\mu}(\mu) = [\psi^{'}(\mu)]^2 I_{\theta}(\psi(\mu))$.
- Пусть $X_i \sim \mathrm{Bin}(10, \theta)$, то есть $p_{x|\theta} = C_{10}^{x}\theta^{x}(1-\theta)^{10-x}$. Найдите $\hat{\theta}_{ML}$, если $\sum\limits_{i}^{100} Xi = 70$.
- Проверьте гипотезу $H_0 : \theta^3 = 0.03$ против $H_1 : \theta^3 ! = 0.03$ при помощи теста Вальда на уровне значимости 5%

Решение

- *TBA*
- Выпишем функцию правдоподобия в явном виде:

$$L(x|\theta) = \prod_{i=1}^{100} (\theta^{x_i}(1-\theta)^{10-x_i})$$

  Прологарифмируем её:

$$l(x|\theta) = \sum (x_i \ln\theta + (10 - x_i)\ln(1-\theta))$$

-

2.4 LR,LM,W и точечный ММП (ДЗ1 2021, №1)

Условие задачи

Компания "Напиши-ка" производит три вида ручек: синие, красные и зелёные. Аналитик компании, Данил, хочет понять, какая ручка выстрелит, а какая не будет пользоваться популярностью. Он анализирует выборку из 300 проданных ручек, среди которых оказалось 150 синих, 100 красных и 50 зелёных. Данил уверен, что ручки продаются независимо друг от друга и вероятность того, что будет продана синяя, он обозначает за p_1, а что будет продана красная – за p_2.

1. Обозначим $p = [p_1 \quad p_2]^\top$, найдите $\hat{p}_{ML}$, оценку максимального правдоподобия.

2. Проверьте гипотезу

$$\begin{cases} H_0 \ : \ p_1 = 0.2, \\ H_A \ : \ p_1 \neq 0.2 \end{cases}$$

   на уровне значимости 0.05 с помощью тестов LR и LM.

3. Проверьте гипотезу

$$\begin{cases} H_0 \ : \ p = \begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix}, \\ H_A \ : \ p \neq \begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix} \end{cases}$$

   на уровне значимости 0.05 с помощью тестов LR и W.