ZK 2020

pondělí 16. ledna 2023 11:06

1. (5 points) Describe the K-means algorithm (1pt). Prove that K-means converges to a local minimum (2pts). Describe the K-means++ algorithm (1pt). Give an example when employing K-means++ leads, on average, to better solution than K-means (1pt).

2. (5 points) Formulate the logistic regression problem (1 point). Derive the gradient descent learning method for this task (2.5 points).
Given the training set $T = \{(-1,-1),(-2,-1),(0,1),(2,1)\}$, starting point $w = (0,0)$, and an initial step size of 1, do the first iteration of gradient descent. (1.5 points).

$$g(\mathbf{w}) = \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \sum_{(\mathbf{x},k)\in\mathcal{T}} \frac{e^{-k\mathbf{w}\cdot\mathbf{x}}}{1+e^{-k\mathbf{w}\cdot\mathbf{x}}}(-k\mathbf{x}) = -\sum_{(\mathbf{x},k)\in\mathcal{T}} \underbrace{\frac{1}{1+e^{k\mathbf{w}\cdot\mathbf{x}}}}_{p(-k|\mathbf{x})} k\mathbf{x}$$

$$= -\sum_{(\mathbf{x},k)\in\mathcal{T}} (1-p(k|\mathbf{x}))\, k\mathbf{x}\,.$$

$$E(\mathbf{w}) = -\sum_{(\mathbf{x},k)\in\mathcal{T}} \ln p(k|\mathbf{x}) = \sum_{(\mathbf{x},k)\in\mathcal{T}} \ln(1+e^{-k\mathbf{w}\cdot\mathbf{x}})$$

$$\underbrace{a(x) = \ln\frac{p(1|x)}{p(2|x)}}_{\text{log odds}} = \vec{w}\cdot\vec{x} \rightarrow \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_D \end{bmatrix} \longrightarrow p(1|x) = \frac{1}{1+e^{-kwx}}$$

$$[\underset{\text{bias}}{w_0}, w_1, \dots, w_D] \qquad p(2|x) = \frac{1}{1+e^{+kwx}}$$

gradient descent
– chci aby E klesalo

$$E(w) = -\ell'(w) = -\sum \ln(p(k|x)) = -\sum \ln(1+e^{-kwx})$$

$$g(w) = \frac{\partial E(w)}{\partial w} = -\sum \frac{1}{(1+e^{kwx})}\cdot kx = -\sum(1-p(k/x))\cdot kx$$

– chci aby $g(w)=0$

---

gradient descent iteration $T = \{(\overset{x}{-1},\overset{k}{-1}),(-2,-1),(0,1),(2,1)\}$

$$\boxed{\begin{aligned} E(w) &- \sum \ln(1+e^{-kwx}) \\ g(w) &= \sum(1-p(k/x))\cdot kx \end{aligned}} - \sum \frac{1}{1+e^{kwx}}\cdot kx$$

$w = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $x_i = (1, x)$
$-k = \pm 1$
step = 1 $1\cdot w\cdot x_i$

$$E(w_0) = \ln(1+\underbrace{e}_{e^{\overbrace{[0,0]\cdot\begin{bmatrix}1\\-1\end{bmatrix}}} = e^0 = 1}) + \ln(1+\underbrace{e^{1\cdot[0]\cdot[1,-2]}}_{e^0}) + \underbrace{\dots}_{\dots e^0 \dots} = \ln(2)+\ln(2)+\dots = \ln(2)\cdot 4$$

$[0,0]\cdot\begin{bmatrix}1\\-1\end{bmatrix} = 0+0 = 0$

$$g(w_0) = \underbrace{\frac{(-1)\cdot[1,-1]}{1+\underbrace{e^{1)\cdot[0]\,[1,-1]}}_{e^0=1}}}_{\frac{(-1,1)}{2}} + \dots = \frac{1}{2}\cdot\Big((-1,1)+(-1,2)+(-1,0)+(-1,-2)\Big) = \frac{1}{2}\cdot(-4,1) = (-2, 1/2)$$

iterace $t=1$

$$w_1 = w - \text{step}\cdot g = (0,0) - 1\cdot(-2, 1/2) = (2, -1/2)$$

$$E_1(w) = \sum \ln(1+e^{-kwx})$$

$x_i = (1,-1,-1)$

$$= \underbrace{\ln(1+e^{\overbrace{(2,-1/2]\cdot\begin{bmatrix}+1\\-1\end{bmatrix}}^{=2+1=3}})}_{\ln(1+e^3)}$$

3. (5 points) Parameter Estimation. The reliability of light bulbs is defined by the probability density function

$$p(t) = \frac{1}{\theta\left(\frac{t}{\theta}+1\right)^2}, \quad t\in[0,\infty]\,,$$

where $p(t)\delta$ is the probability that a bulb fails in a short time interval $(t, t+\delta)$. That is, the probability $P(T)$ that the bulb fails in interval $(0,T)$, is $P(T) = \int_0^T p(t)\mathrm{d}t = 1 - \frac{1}{(\frac{T}{\theta}+1)}$. What is the maximum likelihood estimate of the reliability
a) parameter θ if in an experiment with two bulbs the following lifetimes have been observed: $t_1 = 1.0, t_2 = 4.0$? (4pts)
b) For this setting, what is the probability that a light bulb is still OK at time 10? (1pt)

$$\text{ML}: \hat{\theta} = \arg\max L(\theta) = \arg\max \underbrace{P(T/\theta)}_{\prod_{i=1}^{N} p(x_i,\theta)}$$

$$L(\theta) = \frac{1}{\theta\left(\frac{t_1}{\theta}+1\right)^2}\cdot\frac{1}{\theta\left(\frac{t_2}{\theta}+1\right)^2}$$

$$\ell(\theta) = \ln(L(\theta)) = \ln\frac{1}{\theta\left(\frac{t_1}{\theta}+1\right)^2} + \ln\frac{1}{\theta\left(\frac{t_2}{\theta}+1\right)^2}$$

$$\ell(\theta) = \underbrace{\ln(1)}_{0} - \ln\left(\theta\cdot\left(\frac{t_1}{\theta_1}+1\right)^2\right) + \underbrace{\ln(1)}_{0} - \ln(\dots)$$

$$\ell(\theta) = -\ln(\theta) - \ln\left(\frac{t_1}{\theta}+1\right)^2 - \ln(\theta) - \ln\left(\frac{t_2}{\theta}+1\right)^2$$

$$\ell(\theta) = -\cancel{2\ln(\theta)} - \underbrace{\quad}_{\frac{1}{t_1+\theta}} \quad -2\ln(t_2+\theta) + 2\ln(\theta)$$

$$1 - \frac{1}{\left(\frac{T}{\theta}+1\right)}$$

$$b)\ P(b) = 1 - P(10) = +\frac{1}{\frac{T}{\theta}+1}$$

$$= +\frac{1}{\frac{10}{2}+5} = +\frac{1}{5+5} = +\frac{1}{10} = \underline{\underline{0.1}}$$

$$\ell(\theta) = -\overset{\vee}{\ln}(\theta) - \ln\left(\frac{t_1}{\theta}+1\right)^2 - \overset{0}{\ln}(\theta) - \ln\left(\frac{t_2}{\theta}+1\right)^2$$

$$\ell(\theta) = -\cancel{2\ln}(\theta) - \underbrace{\underbrace{\underbrace{\frac{t_1}{\theta}+1}_{\frac{t_1+\theta}{\theta}}}_{-2\ln\left(\frac{t_1+\theta}{\theta}\right)}}_{-2\ln(t_1+\theta)+\cancel{2\ln}(\theta)} - 2\ln(t_2+\theta) + 2\ln(\theta)$$

$$\ell(\theta) = 2\ln(\theta) - 2\ln(t_1+\theta) - 2\ln(t_2+\theta)$$

$$\frac{\partial \ell(\theta)}{\partial\theta} = \frac{2}{\theta} - \frac{2}{t_1+\theta} - \frac{2}{t_2+\theta} = 0$$

$$\left.\begin{array}{l} t_1 = 1.0 \\ t_2 = 4.0 \end{array}\right\} \quad \frac{2}{\theta} - \frac{2}{1+\theta} - \frac{2}{4+\theta} = 0$$

$$\frac{2\cdot(1+\theta)\cdot(4+\theta) - 2\theta\cdot(4+\theta) - 2\theta\cdot(1+\theta)}{\theta\cdot(1+\theta)\cdot(4+\theta)} = 0$$

$$(2+2\theta)\cdot(4+\theta) - 8\theta - 2\theta^2 - 2\theta - 2\theta^2 = 0$$

$$8 + \cancel{2\theta} + \cancel{8\theta} + \cancel{2\theta^2} - \cancel{8\theta} - \cancel{2\theta^2} - \cancel{2\theta} - 2\theta^2 = 0$$

$$-2\theta^2 + 8 = 0$$

$$\theta^2 = \frac{-8}{-2} = 4$$

$$\theta = \pm\sqrt{4} = \underline{\underline{\pm 2}}$$

4. (5 points) Decision tree classifiers.

   (a) (2 points) Describe the simplest top-down tree-building algorithm that selects the decision rule at a node by maximization of information gain.

   (b) (3 points) Discuss advantages and disadvantages of decision trees.