CS 229, Autumn 2016
Problem Set #3 Solutions: Theory & Unsupervised learning

Qingxin6174

1. (a) $\psi_i = P(|\hat{\phi}_i - \phi_i| > \gamma) \leqslant 2e^{-2\gamma^2 m}$.

   (b) *Proof.* Let $\psi_{V_i} = P(V_i = 1) \leqslant P(W_i = 1) = \psi_{W_i}$, $\forall i = \{1, 2, \cdots, k\}$.

   If $t < 0$, $P\left(\sum_{i=1}^{k} V_i > t\right) = 1 = P\left(\sum_{i=1}^{k} W_i > t\right)$, if $t \geqslant k$, $P\left(\sum_{i=1}^{k} V_i > t\right) = 0 = P\left(\sum_{i=1}^{k} W_i > t\right)$,

   if $t = 0$, $P\left(\sum_{i=1}^{k} V_i > 0\right) = 1 - \prod_{i=1}^{k}(1-\psi_{V_i}) \leqslant 1 - \prod_{i=1}^{k}(1-\psi_{W_i}) = P\left(\sum_{i=1}^{k} W_i > 0\right)$,

   if $0 < t < k$, let $z = \min_{\mathrm{z}}\{\, \mathrm{z} \in \mathbb{Z} \mid \mathrm{z} > t \,\}$, $\psi_{V_l} \in \{\psi_{V_1}, \psi_{V_2}, \cdots, \psi_{V_k}\}$, and let

   $$f(\psi_{V_1}, \psi_{V_2}, \cdots, \psi_{V_k}) = \sum_{j=z}^{k} P\left(\sum_{i=1}^{k} V_i = j\right) = \sum_{\sum_{i=1}^{k} 1\{V_i=1\}=z}^{k} \prod_{i=1}^{k} \psi_{V_i}^{1\{V_i=1\}}(1-\psi_{V_i})^{1\{V_i=0\}},$$

   $$\begin{aligned}\frac{\partial f}{\partial \psi_{V_l}} &= \frac{\partial}{\partial \psi_{V_l}} \sum_{\sum_{i=1}^{k} 1\{V_i=1\}=z}^{k} \psi_{V_l}^{1\{V_l=1\}}(1-\psi_{V_l})^{1\{V_l=0\}} \prod_{i=1,\, i\neq l}^{k} \psi_{V_i}^{1\{V_i=1\}}(1-\psi_{V_i})^{1\{V_i=0\}} \\ &= \frac{\partial}{\partial \psi_{V_l}} \left(\sum_{\sum_{i=1,\, i\neq l}^{k} 1\{V_i=1\}=z-1}^{k-1} \psi_{V_l} \prod_{i=1,\, i\neq l}^{k} \psi_{V_i}^{1\{V_i=1\}}(1-\psi_{V_i})^{1\{V_i=0\}} \right. \\ &\qquad \left. + \sum_{\sum_{i=1,\, i\neq l}^{k} 1\{V_i=1\}=z}^{k-1} (1-\psi_{V_l}) \prod_{i=1,\, i\neq l}^{k} \psi_{V_i}^{1\{V_i=1\}}(1-\psi_{V_i})^{1\{V_i=0\}} \right) \\ &= \sum_{\sum_{i=1,\, i\neq l}^{k} 1\{V_i=1\}=z-1} \prod_{i=1,\, i\neq l}^{k} \psi_{V_i}^{1\{V_i=1\}}(1-\psi_{V_i})^{1\{V_i=0\}} \geqslant 0,\end{aligned}$$

   then $P\left(\sum_{i=1}^{k} V_i > t\right) = f(\psi_{V_1}, \psi_{V_2}, \cdots, \psi_{V_k}) \leqslant f(\psi_{W_1}, \psi_{W_2}, \cdots, \psi_{W_k}) = P\left(\sum_{i=1}^{k} W_i > t\right)$.

   From the above: $\forall t \in \mathbb{R}$, $P\left(\sum_{i=1}^{k} V_i > t\right) \leqslant P\left(\sum_{i=1}^{k} W_i > t\right)$. $\square$

   (c) *Proof.* $Z_i \sim$ Brenoulli(ψ_i), and let $W_i \sim$ Brenoulli(p), where $p = 2e^{-2\gamma^2 m}$. Consider $\tau < 1$,

   $$P\left(\overline{Z} > \tau\right) = P\left(\sum_{i=1}^{n} Z_i > n\tau\right) \leqslant P\left(\sum_{i=1}^{n} W_i > n\tau\right) \leqslant e^{-nD(\tau||p)}, \qquad (\exists m, \text{ s.t. } p < \tau)$$

   where $D(\tau||p) = \tau \log \dfrac{\tau}{p} + (1-\tau)\log\dfrac{1-\tau}{1-p}$. $\square$

2. *Proof.* $\forall \theta \in \mathbb{R}^{d+1}$, $\left|\left\{ x \,\middle|\, \sum_{i=0}^{d} \theta_i x^i = 0,\ x \in \mathbb{R} \right\}\right| \leqslant d$. $\forall \{(p_j, y_i) \mid p_j \in \mathbb{R}, y_j \in \{-1, +1\}\}_{j=1}^{t+1}$,

   $k \neq l \Rightarrow p_k \neq p_l$, $y_j y_{j+1} < 0 \Rightarrow \exists x_j' \in (p_j, p_{j+1}) \Rightarrow |\{x_j'\}| \leqslant t$, let $\{x_j'\} \subseteq \left\{ x \,\middle|\, \sum_{i=0}^{d} \theta_i x^i = 0,\ x \in \mathbb{R} \right\}$,

   then $t \leqslant d$, $VC(\mathcal{H}) = \sup t + 1 = d + 1$. $\square$

3. *Proof.* $\theta \sim \mathcal{N}(0,\, \tau^2 I)$, $p(\theta;\, 0,\, \tau^2 I) = \dfrac{1}{(2\pi)^{\frac{n}{2}} |\tau^2 I|^{\frac{1}{2}}} e^{-\frac{1}{2}\theta^T \left(\tau^2 I\right)^{-1}\theta} = (2\pi\tau^2)^{-\frac{n}{2}} e^{-\frac{1}{2\tau^2}\|\theta\|_2^2}$.

$$\begin{aligned}
\theta_{\mathrm{ML}} &= \arg\max_\theta \prod_{i=1}^m \left(h_\theta\left(x^{(i)}\right)\right)^{1\{y^{(i)}=1\}} \left(1 - h_\theta\left(x^{(i)}\right)\right)^{1\{y^{(i)}=0\}} \\
&= \arg\max_\theta \sum_{i=1}^m \left[1\{y^{(i)} = 1\} \log h_\theta\left(x^{(i)}\right) + 1\{y^{(i)} = 0\} \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right], \\
\theta_{\mathrm{MAP}} &= \arg\max_\theta (2\pi\tau^2)^{-\frac{n}{2}} e^{-\frac{1}{2\tau^2}\|\theta\|_2^2} \prod_{i=1}^m \left(h_\theta\left(x^{(i)}\right)\right)^{1\{y^{(i)}=1\}} \left(1 - h_\theta\left(x^{(i)}\right)\right)^{1\{y^{(i)}=0\}} \\
&= \arg\max_\theta \sum_{i=1}^m \left[1\{y^{(i)} = 1\} \log h_\theta\left(x^{(i)}\right) + 1\{y^{(i)} = 0\} \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] - \frac{1}{2\tau^2}\|\theta\|_2^2.
\end{aligned}$$

Assume that $\|\theta_{\mathrm{MAP}}\|_2 > \|\theta_{\mathrm{ML}}\|_2$,

$$\begin{aligned}
&\sum_{i=1}^m \left[1\{y^{(i)} = 1\} \log h_{\theta_{\mathrm{MAP}}}\left(x^{(i)}\right) + 1\{y^{(i)} = 0\} \log\left(1 - h_{\theta_{\mathrm{MAP}}}\left(x^{(i)}\right)\right)\right] - \frac{1}{2\tau^2}\|\theta_{\mathrm{MAP}}\|_2^2 \\
< &\sum_{i=1}^m \left[1\{y^{(i)} = 1\} \log h_{\theta_{\mathrm{ML}}}\left(x^{(i)}\right) + 1\{y^{(i)} = 0\} \log\left(1 - h_{\theta_{\mathrm{ML}}}\left(x^{(i)}\right)\right)\right] - \frac{1}{2\tau^2}\|\theta_{\mathrm{ML}}\|_2^2 \Rightarrow \text{contradiction},
\end{aligned}$$

this implies that $\|\theta_{\mathrm{MAP}}\|_2 \leqslant \|\theta_{\mathrm{ML}}\|_2$. □

4. (a) *Proof.* $KL(P\|Q) = -\sum_x P(x) \log \dfrac{Q(x)}{P(x)} \geqslant -\sum_x P(x) \left(\dfrac{Q(x)}{P(x)} - 1\right) = \sum_x P(x) - \sum_x Q(x) = 0$,

   for equality to hold, $\log \dfrac{Q(x)}{P(x)} = \dfrac{Q(x)}{P(x)} - 1$, which can happen if and only if $P = Q$. □

   (b) *Proof.* Start on the right side:

$$\begin{aligned}
& KL(P(X)\|Q(X)) + KL(P(Y|X)\|Q(Y|X)) \\
= & \left(\sum_x P(x) \log \frac{P(x)}{Q(x)}\right) \left(\sum_y P(y|x)\right) + \sum_x P(x) \left(\sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)}\right) \\
= & \sum_{x,\, y} P(x,\, y) \log \frac{P(x)}{Q(x)} + \sum_{x,\, y} P(x,\, y) \log \frac{P(y|x)}{Q(y|x)} \\
= & \sum_{x,\, y} P(x,\, y) \log \frac{P(x,\, y)}{Q(x,\, y)} \\
= & KL(P(X,\, Y)\|Q(X,\, Y)),
\end{aligned}$$

   chain rule holds true. □

   (c) *Proof.* Expand the left side by the definition of KL divergence,

$$\begin{aligned}
\arg\min_\theta KL(\hat{P}\|P_\theta) &= \arg\min_\theta \sum_x \hat{P}(x) \log \hat{P}(x) - \sum_x \hat{P}(x) \log P_\theta(x) \\
&= \arg\max_\theta \sum_x \hat{P}(x) \log P_\theta(x) \\
&= \arg\max_\theta \sum_x \frac{1}{m} \sum_{i=1}^m 1\left\{x^{(i)} = x\right\} \log P_\theta(x) \\
&= \arg\max_\theta \sum_{i=1}^m \sum_x 1\left\{x^{(i)} = x\right\} \log P_\theta(x) \\
&= \arg\max_\theta \sum_{i=1}^m \log P_\theta\left(x^{(i)}\right).
\end{aligned}$$

   Finding maximum likelihood is equivalent to finding minimal KL divergence from $\hat{P}$. □

5. My implementation of `K_means_img_compress.m`

```
function A_compressed = K_means_img_compress (A, k)
[m,n,p] = size(A); Points = reshape(A,[m*n,p]);
centroid = Points(randperm(m*n,k),:); idx_old = zeros(m*n,1); err = 2;
while err >= 1
    [~,idx] = min(pdist2(Points,centroid),[],2);
    for i = 1:k
        centroid(i,:) = round(mean(Points(idx==i,:),1));
    end
    err = sum(abs(idx-idx_old)); idx_old = idx;
end
A_compressed = reshape(centroid(idx,:),m,n,p);
```

Figure 1 shows the cluster state in pixel's space using *mandrill-small.tiff* when $k = 16$, figure 2 shows the *mandrill-large.tiff* and its' compressed image.

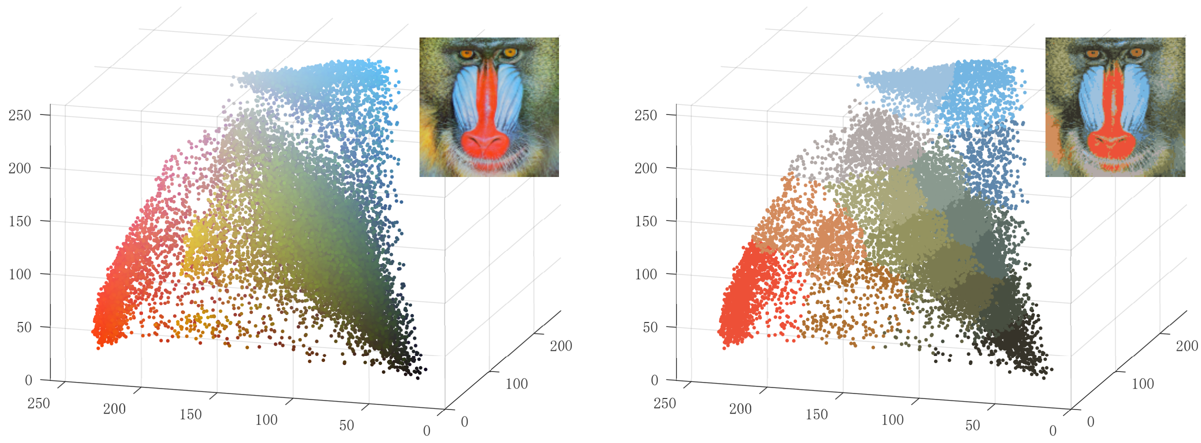

Fig. 1: 128×128 pixels, original image vs. compressed image, and cluster in pixel's space.

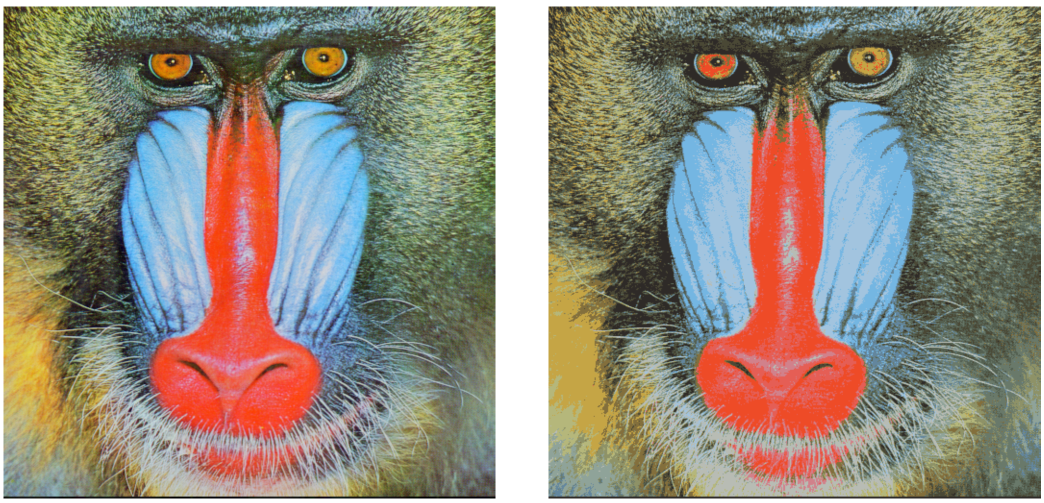

Fig. 2: 512×512 pixels, original image vs. compressed image.

The size of the original image is $512 \times 512 \times 24$ bits $= 786{,}432$ bytes, the size of the compressed image is $512 \times 512 \times 4$ bits index $+ 16 \times 24$ bits $= 131{,}120$ bytes.