CSE517 HW4 Ziyang Jiao 475589
Flora Sun 475178

Problem 1.

For noise free observations:

$$\text{cov}(f_*) = k_{**} - k_*^T K^{-1} k_*$$

As x_* is a training point. Let's assume $x^* = x_i \quad i \in \{1, \dots, n\}$

Then $k_* = Ke_i$, $k_{**} = e_i^T K e_i$

where e_i is a column vector with the i-th element be 1 and others be 0.

$$\therefore \text{cov}(f_*) = e_i^T K e_i - (Ke_i)^T K^{-1} (Ke_i)$$
$$= e_i^T K e_i - e_i^T K^T K^{-1} K e_i$$
$$= e_i^T K e_i - e_i^T K^T e_i$$

$\because$ K is the covariance matrix of X, K is symmetric.
↳ (gram matrix if we use kernel)

$$\therefore K^T = K$$

$$\therefore \text{cov}(f_*) = e_i^T K e_i - e_i^T K e_i$$
$$= 0$$

Problem 2.

a) $p(\vec{y} \mid x, \vec{\theta}) = p(\vec{f} + \varepsilon \mid x)$

$$= N(0, K_\theta(x,x) + \sigma^2 I)$$

$$\therefore \log p(\vec{y} \mid x, \vec{\theta}) = \log \left(\frac{1}{\sqrt{(2\pi)^n |K_\theta(x,x) + \sigma^2 I|}} e^{-\frac{1}{2}\vec{y}^T (K_\theta(x,x) + \sigma^2 I)^{-1} \vec{y}} \right)$$

$$= -\frac{1}{2}\log\left[(2\pi)^n \cdot |K_\theta(x,x) + \sigma^2 I|\right] - \frac{1}{2}\vec{y}^T (K_\theta(x,x) + \sigma^2 I)^{-1} \vec{y}$$

$$= -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\left(|K_\theta(x,x) + \sigma^2 I|\right) - \frac{1}{2}\vec{y}^T (K_\theta(x,x) + \sigma^2 I)^{-1} \vec{y}$$

Ziyang Jiao 47558
Flora Sun 47518

b) $\because \log(\vec{y}|x,\vec{\theta}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log(|K_y|) - \frac{1}{2}\vec{y}^T K_y^{-1}\vec{y}$

$= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log(|LL^T|) - \frac{1}{2}\vec{y}^T (LL^T)^{-1}\vec{y}$

$= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log|LL^T| - \frac{1}{2}\vec{y}^T L^{-T}L^{-1}\vec{y}$

$\because \alpha = L^T \backslash (L \backslash y)$

$\therefore L^T\alpha = L \backslash y$

$\therefore LL^T\vec{\alpha} = \vec{y}$

$\therefore \log(\vec{y}|x,\vec{\theta}) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log|LL^T| - \frac{1}{2}\vec{\alpha}^T LL^T (L^T)^{-1}L^{-1}LL^T\vec{\alpha}$

$= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log|LL^T| - \frac{1}{2}\vec{\alpha}^T LL^T\vec{\alpha}$

$= -\frac{1}{2}\log 2\pi - \sum_{i=1}^{n}\log L_{ii} - \frac{1}{2}\vec{\alpha}^T LL^T\vec{\alpha}$

c) $\frac{\partial \log(\vec{y}|x,\vec{\theta})}{\partial \vec{\theta}} = \frac{\partial \log(\vec{y}|x,\vec{\theta})}{\partial K_y} \cdot \frac{\partial K_y}{\partial \theta}$

$= -\frac{n}{2}\log 2\pi - \frac{1}{2}\log(|K_y|) - \frac{1}{2}\vec{y}^T K_y^{-1}\vec{y}^T$

$= -\frac{1}{2}\frac{|K_y|K_y^{-1}}{|K_y|}\cdot\frac{\partial K_y}{\partial\theta} - \frac{1}{2}\vec{y}^T K_y^{-1}\frac{\partial K_y}{\partial\theta}K_y^{-1}\vec{y}$

$= -\frac{1}{2}K_y^{-1}\frac{\partial K_y}{\partial\theta} - \frac{1}{2}\vec{y}^T K_y^{-1}\frac{\partial K_y}{\partial\theta}K_y^{-1}\vec{y}$

$= -\frac{1}{2}(LL^T)^{-1}\frac{\partial K_y}{\partial\theta} - \frac{1}{2}\vec{y}^T(LL^T)^{-1}\frac{\partial K_y}{\partial\theta}(LL^T)^{-1}\vec{y}$

$= -\frac{1}{2}(L^{-1})^T L^{-1}\frac{\partial K_y}{\partial\theta} - \frac{1}{2}\vec{\alpha}^T\frac{\partial K_y}{\partial\theta}\vec{\alpha}$

Ziyang Jiao 475589
Flora Sun 475178

Problem 3:

(a) (i) ⇔ (ii)

prove (i) ⇒ (ii):

Assume in i-th iteration:

we got Assignment A.

Then we will update the center with:

$$\mu_\alpha = \frac{\sum_{i=1}^{n} [z_i]_\alpha x_i}{\sum_{i=1}^{n} [z_i]_\alpha}$$

AKA the mean vector of datas that assigned to the same cluster.

Then in (i+1)-th iteration:

if we still got Assignment A

Then μ_α will also be the same: $\mu_\alpha = \frac{\sum_{i=1}^{n} [z_i]_\alpha x_i}{\sum_{i=1}^{n} [z_i]_\alpha}$

∴ (i) ⇒ (ii)

prove (ii) ⇒ (i)

Assume in i th iteration:

we got Assignment A

Then we update the center to be $\mu_1, \ldots, \mu_k$.

Then in (i+1)-th iteration:

we got Assignment B (assign according to the closest center)

Then we update the center, but the center stay the same.

Then in the (i+2)th iteration.

As we still assign data according to the closest center to it. we will end up with the same assignment B.

∴ (ii) ⇒ (i)

In conclusion:

(i) ⇔ (ii)

b) The k-means will always converge:

As there is only k^n different assignments on n data points. As we want to minimize the "variance" inside each cluster. There must be a infimum bound of all this assignments. As we upate the cluster each time to achieve lower "variance" inside each cluster. And the intimum is finite. Then we must will converge.

c) It is possible that k means generate empty cluster. for example when we initial the init_center into this:

However we should not allow it.

Because if k-means generate empty cluster, we can always fetch a data point from any other not-empty cluster and make it as this empty cluster center. Then we will achieve lower in cluster variance and achieve better result.

Ziyang Jiao 475589
Flora Sun 475178

Problem 4.

(Reference: 12 lecture note. kernel k means)

a) Randomly initialize $z_1, \ldots, z_n$

repeat
  for $i = 1, \ldots, n$ do
    for $\alpha = 1, \ldots, k$ do
      $$dia \leftarrow k(x_i, x_i) - \frac{2}{\sum_{j=1}^{n} [z_j]_\alpha} \sum_{j=1}^{n} [z_j] k(x_i, x_j) + \frac{1}{(\sum_{j=1}^{n} [z_j]_\alpha)^2} \sum_{j=1}^{n} \sum_{l=1}^{n} [z_j]_\alpha [z_l]_\alpha k(x_j, x_l)$$
    end for
  end for
  update z_i using $\alpha = \text{argmin}_\alpha dia$
  if z_i do not chang for all i. then
    STOP
  end if
until convergence.

b) Because the cluster center "lies" in the transformed feature space. For some kernel like RBF, it has infinite dimensions. Thus is impossible to visualize or use them as prototypes for learning.

CSE 517 HW4 Ziyang Jiao 475589
Flora Sun 475178

Problem 6

a) As $\|\vec{x} - \vec{u}\vec{u}^T\vec{x}\|^2 (\text{s.t. } u^Tu=1)$ is convex and differentiable

∴ local minimizer is the global minimizer. (convex)

and global minimizer will satisfy: $\frac{\partial \|\vec{x} - \vec{u}\vec{u}^T\vec{x}\|^2 - \lambda(\vec{u}^T\vec{u}-1)}{\partial \vec{u}} = \vec{0}$

$$\|\vec{x} - uu^T\vec{x}\|^2$$
$$= (\vec{x} - uu^T\vec{x})^T(\vec{x} - uu^T\vec{x})$$
$$= \vec{x}^T\vec{x} - 2\vec{x}^Tuu^T\vec{x} + \vec{x}^Tuu^Tuu^T\vec{x} \quad (\vec{u} \text{ is orthogonal})$$
$$= \vec{x}^T\vec{x} - \vec{u}^T\vec{x}\vec{x}^T\vec{u}$$

$$\frac{\partial \|\vec{x} - \vec{u}\vec{u}^T\vec{x}\|^2 - \lambda(\vec{u}^T\vec{u}-1)}{\partial \vec{u}} = -2\vec{x}\vec{x}^T\vec{u} - 2\lambda\vec{u}$$

In PCA, $u = u_e$ when u_e is the eigen vector matrix of C.

$$C = \vec{x}\vec{x}^T$$

∴ $-2\vec{x}\vec{x}^T\vec{u} - 2\lambda\vec{u} = -2C\vec{u} - 2\lambda\vec{u}$

∵ $C\vec{u} = \lambda\vec{u}$

∴ $\frac{\partial \|\vec{x} - \vec{u}\vec{u}^T\vec{x}\|^2 - \lambda(\vec{u}^T\vec{u}-1)}{\partial \vec{u}} = -2(C\vec{u} - \lambda\vec{u}) = \vec{0}$

∴ PCA solution is optimal.

b) PCA feature transformation $\tilde{X} = u^TX$ where u is the eigen-vector matrix of XX^T.

Normalization: $X = \frac{X - \mu(X)}{std(X)}$

scaling: $X = \frac{X - X_{min}}{X_{max} - X_{min}}$

Ziyang Jiao 475589
Flora Sun 475178

After PCA, each feature are uncorrelated with each other. This is not true for either normalization nor scaling. PCA can be used as dimensional reduction method, while normalization and scaling can not.

After normalization: $u=0$ and $\sigma=1$. This might not be true for scaling. After normalization, the relative mahalanobis distance between two points do not change. While after scaling. the relative euclidean distance between two points do not change.

Scaling works better when the data do not obey gaussian. After scaling the data lies in $[0, 1]$ which is not true for normalization

(c) ① if we use gradient method to find optimal weights for both logistic regression and ridge regression, It will converge faster.

② with feature scaling / feature normalization, we can avoid that large-scale feature matter more than those small-scale features. (Because the regularization might limit the weights in front of small-scale features, so they can not contribute to the prediction as much as those large-scale features.

Problem 5:

a) $\pi_j \cdot \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} \exp(-\frac{1}{2}(\vec{x}-\vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}-\vec{\mu}_j))$

b) $L(x|\theta) = \prod_{i=1}^{n} (\sum_{j=1}^{k} \pi_j p(\vec{x}_i|\theta_j))$

if we put log over it: $\Rightarrow \sum_{i=1}^{n} \log \sum_{j=1}^{k} \pi_j p(\vec{x}_i|\theta_j)$

c) We could plot GMM against k. And use the k after which GMM likelihood stop increasing rapidly.

d) The log-likelihood is:

$$\log(p(x_i|\theta_j)) = \log \prod_{m=1}^{d} (\theta_{jm})^{x_{im}} (1-\theta_{jm})^{(1-x_{im})}$$
$$= \sum_{m=1}^{d} (x_{im} \log \theta_{jm} + (1-x_{im}) \log(1-\theta_{jm}))$$

maximize $B = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \log(\frac{\pi_j p(x_i|\vec{\theta}_j)}{z_{ij}})$

maximize $\sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} (\log \pi_j + \sum_{m=1}^{d} (x_{im} \log \theta_{jm} + (1-x_{im}) \log(1-\theta_{jm})) - \log z_{ij})$

$$\frac{\partial}{\partial \theta_{jm}} \left(\sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} (\log \pi_j + \sum_{m=1}^{d} (x_{im} \log \theta_{jm} + (1-x_{im}) \log(1-\theta_{jm})) - \log z_{ij}) \right)$$
$$= \sum_{i=1}^{n} z_{ij} \left(\frac{x_{im}}{\theta_{jm}} - \frac{1-x_{im}}{1-\theta_{jm}} \right)$$

to maximize this: set the derivative to be 0:

$$\sum_{i=1}^{n} z_{ij} \left(\frac{x_{im}}{\theta_{jm}} - \frac{1-x_{im}}{1-\theta_{jm}} \right) = 0$$

$$\sum_{i=1}^{n} z_{ij} \left(\frac{x_{im} - \theta_{jm}}{\theta_{jm}(1-\theta_{jm})} \right) = 0$$

Ziyang Jiao 475589
Flora Sun 475178

$$\therefore \sum_{i=1}^{n} z_{ij}(x_{im} - \theta_{jm}) = 0 \qquad (0 < \theta_{jm} < 1)$$

$$\therefore \theta_{jm} = \frac{\sum_{i=1}^{n} z_{ij} x_{im}}{\sum_{i=1}^{m} z_{ij}}$$