ML1 Home Assignment 4

id: 12179078

Volodymyr Medentsiy

w1

K - experts, N - datapoints $\{(x_i, y_i)\}_{i=1}^N$ with $x_i \in \mathbb{R}^D$

y_i - label of x_i; z_n - 1-of-K vector.

$\theta \in \mathbb{R}^{D \times K}$ - vector of parameters for each expert.

$$p(y_n \mid x_n, z_n, \theta) = \text{Exp}(y_n \mid \lambda = \exp(\theta_k^T x_n))$$

$$p(z_n = k \mid x_n, \varphi) = \pi_{nk} = \frac{\exp(\varphi_k^T x_n)}{\sum_j \exp(\varphi_j^T x_n)}$$

$\varphi \in \mathbb{R}^{D \times K}$ - parameters of the routing mechanism.

($z_n = k$ means that $z_{nk} = 1$).

I Part

1.1. The likelihood of the entire dataset $p(y \mid X, \varphi, \theta)$ and its log under i.i.d assumption.

$$p(y \mid X, \varphi, \theta) = \left[\text{i.i.d. assumption}\right] = \prod_{n=1}^N p(y_n \mid x_n, \theta, \varphi) =$$

$$= \prod_{n=1}^N \sum_{k=1}^K p(z_n = k \mid x_n, \varphi) \cdot p(y_n \mid x_n, z_n, \theta) =$$

$$= \prod_{n=1}^N \sum_{k=1}^K \frac{\exp(\varphi_k^T x_n)}{\sum_{j=1}^K \exp(\varphi_j^T x_n)} \cdot \text{Exp}(y_n \mid \lambda = \exp(\theta_k^T x_n)) =$$

$$= \prod_{n=1}^N \sum_{k=1}^K \pi_{nk} \text{Exp}(y_n \mid \lambda = \exp(\theta_k^T x_n))$$

$$\log p(y \mid X, \varphi, \theta) = \sum_{n=1}^N \log \sum_{k=1}^K \frac{\exp(\varphi_k^T x_n)}{\sum_{j=1}^K \exp(\varphi_j^T x_n)} \text{Exp}(y_n \mid \lambda = \exp(\theta_k^T x_n)) =$$

$$= \sum_{n=1}^N \log \sum_{k=1}^K \pi_{nk} \text{Exp}(y_n \mid \lambda = \exp(\theta_k^T x_n)).$$

1.2. Find the responsibility (r_{ni}) of the expert i for datapoint n.

$$r_{ni} = p(z_n = i \mid y_n, x_n, \varphi, \theta) = \frac{p(y_n \mid z_n = i, x_n, \theta)\, p(z_n = i \mid x_n, \varphi)}{p(y_n \mid x_n, \theta, \varphi)} =$$

$$= \frac{p(y_n \mid z_n = i, x_n, \theta) \cdot p(z_n = i \mid x_n, \varphi)}{\sum_{j=1}^{K} p(z_n = j \mid x_n, \varphi) \cdot p(y_n \mid z_n = j, x_n, \theta)} =$$

$$= \frac{\pi_{ni}\ \mathrm{Exp}(y_n \mid \lambda = \exp(\theta_i^T x_n))}{\sum_{j=1}^{K} \pi_{nj}\ \mathrm{Exp}(y_n \mid \lambda = \exp(\theta_j^T x_n))}$$

1.3.

$$\frac{d}{d\theta_i} \log p(y \mid X, \varphi, \theta) = \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} p(y_n \mid x_n, z_n = k, \theta)\, p(z_n = k \mid x_n, \varphi)} \cdot$$

$$\cdot \frac{d}{d\theta_i} \sum_{k=1}^{K} p(y_n \mid x_n, z_n = k, \theta) \cdot p(z_n = k \mid x_n, \varphi) =$$

$$= \sum_{n=1}^{N} \frac{r_{ni}}{p(z_n = i \mid x_n, \varphi) \cdot p(y_n \mid x_n, z_n = i, \theta)} \cdot p(z_n = i \mid x_n, \varphi) \cdot$$

$$\frac{d}{d\theta_i} p(y_n \mid x_n, z_n = i, \theta) = \sum_{n=1}^{N} \frac{r_{ni}}{p(y_n \mid x_n, z_n = i, \theta)} \frac{d}{d\theta_i} p(y_n \mid x_n, z_n = i, \theta) =$$

$$= \sum_{n=1}^{N} r_{ni} \frac{d}{d\theta_i} \log(p(y_n \mid x_n, z_n = i, \theta))$$

$$\frac{d}{d\varphi_i}\log p(y|X,\varphi,\theta) = \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} p(y_n|x_n, z_n=k,\theta)\, p(z_n=k|x_n,\varphi)} \cdot$$

$$\cdot \frac{d}{d\varphi_i} \sum_{k=1}^{K} p(y_n|x_n, z_n=k,\theta)\cdot p(z_n=k|x_n,\varphi) =$$

$$= \sum_{n=1}^{N} \frac{r_{ni}}{p(z_n=i|x_n,\varphi)\cdot p(y_n|x_n, z_n=i,\theta)} \cdot \sum_{k=1}^{K} p(y_n|x_n, z_n=k)\frac{d}{d\varphi_i} p(z_n=k|x_n,\varphi)\cdot$$

1.4.

1) $$\frac{d}{d\theta_i}\log p(y|X,\varphi,\theta) = \sum_{n=1}^{N} \frac{r_{ni}}{p(y_n|x_n, z_n=i,\theta)} \frac{d}{d\theta_i} p(y_n|x_n, z_n=i,\theta) =$$

$$= \sum_{n=1}^{N} \frac{r_{ni}}{\exp(\theta_i^T x_n)\exp(-\exp(\theta_i^T x_n) y_n)} \frac{d}{d\theta_i}\exp(\theta_i^T x_n)\exp(-\exp(\theta_i^T x_n) y_n) =$$

$$= \sum_{n=1}^{N} \frac{r_{ni}}{\exp(\theta_i^T x_n)\exp(-\exp(\theta_i^T x_n) y_n)} \cdot \Big(\exp(\theta_i^T x_n)\exp(-\exp(\theta_i^T x_n) y_n)\cdot x_n^T -$$

$$- \exp(\theta_i^T x_n)\cdot\exp(-\exp(\theta_i^T x_n) y_n)\cdot \exp(\theta_i^T x_n)\cdot y_n x_n^T \Big) =$$

$$= \sum_{n=1}^{N} r_{ni}\cdot\left(1 - \exp(\theta_i^T x_n) y_n\right) x_n^T.$$

2) $$\frac{d}{d\varphi_i}\log p(y|X,\varphi,\theta) = \sum_{n=1}^{N} \frac{r_{ni}}{\pi_{ni}\exp(\theta_i^T x_n)\exp(-\exp(\theta_i^T x_n) y_n)} \cdot \sum_{k=1}^{K} \Big(\exp(\theta_k^T x_n)\cdot$$

$$\cdot \exp(-\exp(\theta_k^T x_n) y_n)\cdot \frac{d}{d\varphi_i}\frac{\exp(\varphi_i^T x_n)}{\sum_{j=1}^{K}\exp(\varphi_j^T x_n)} =$$

$$\frac{d}{d\varphi_i}\pi_{nk} \overset{k\neq i}{=} \exp(\varphi_k^T x_n)\,(-1)\cdot\frac{1}{\left(\sum_{j=1}^{K}\exp(\varphi_j^T x_n)\right)^2}\cdot\exp(\varphi_i^T x_n)\cdot x_n^T =$$

$$= -\frac{\exp(\varphi_k^T x_n+\varphi_i^T x_n)}{\left(\sum_{j=1}^{K}\exp(\varphi_j^T x_n)\right)^2}\cdot x_n^T = -\pi_{nk}\cdot\pi_{ni}\,x_n^T$$

$$\frac{d}{d\varphi_i}\pi_{ni} = \frac{\exp(\varphi_i^T x_n)\cdot x_n^T}{\sum_{j=1}^{K}\exp(\varphi_j^T x_n)} - \frac{\exp(\varphi_i^T x_n+\varphi_i^T x_n)}{\left(\sum_{j=1}^{K}\exp(\varphi_j^T x_n)\right)^2}\cdot x_n^T =$$

$$= \frac{\exp(\varphi_i^T x_n)}{\sum_{j=1}^{K}\exp(\varphi_j^T x_n)}\cdot\left(1-\frac{\exp(\varphi_i^T x_n)}{\sum_{j=1}^{K}\exp(\varphi_j^T x_n)}\right)x_n^T = \pi_{ni}(1-\pi_{ni})\,x_n^T$$

$$(=)\sum_{n=1}^{N}\frac{r_{ni}}{\pi_{ni}\exp(\theta_i^T x_n)\exp(-\exp(\theta_i^T x_n)y_n)}\cdot\sum_{k=1}^{K}\Big(\exp(\theta_k^T x_n)\cdot\exp(-\exp(\theta_k^T x_n)y_n)\cdot$$

$$\cdot\,\pi_{ni}\left(\delta_{ik}-\pi_{nk}\right)x_n^T\Big) =$$

$$=\sum_{n=1}^{N} r_{ni}\cdot\sum_{k=1}^{K}\Big(\exp\left((\theta_k^T-\theta_i^T)x_n\right)\cdot\exp\left(-(\exp\theta_k^T x_n-\exp\theta_i^T x_n)y_n\right)\cdot\left(\delta_{ik}-\pi_{nk}\right)\Big)x_n^T$$

with $\delta_{ik}=\begin{cases}0, & i\neq k\\ 1, & i=k\end{cases}$

Or we can also rewrite it as:

$$\sum_{n=1}^{N}\frac{r_{ni}}{\pi_{ni}\exp(\theta_i^T x_n)\exp(-\exp(\theta_i^T x_n)y_n)}\cdot\sum_{k=1}^{K}\Big(\exp(\theta_k^T x_n)\cdot\exp(-\exp(\theta_k^T x_n)y_n)\cdot$$

$$\cdot\,\pi_{ni}\left(\delta_{ik}-\pi_{nk}\right)x_n^T\Big) =$$

$$= \sum_{n=1}^{N} \pi_{ni} \frac{\left(-\sum_{k=1}^{K} \exp(\theta_k^T x_n) \cdot \exp(-\exp(\theta_k^T x_n) y_n) \cdot \pi_{nk} + \left((1-\pi_{ni}) + \pi_{ni}\right) \exp(\theta_i^T x_n) \cdot \right.}{\sum_{j=1}^{N} \pi_{nj} \exp(\theta_j^T x_n) \exp(-\exp(\theta_j^T x_n) y_n)}$$

$$\left. \cdot \exp(-\exp(\theta_i^T x_n) y_n) \right) \cdot x_n^T = \sum_{n=1}^{N} \pi_{ni} \left(-1 + \frac{\exp(\theta_i^T x_n) \cdot \exp(-\exp(\theta_i^T x_n)}{\sum_{j=1}^{N} \pi_j \exp(\theta_j^T x_n) \exp(-\exp(\theta_j^T x_n) y_n)} \right) x_n^T =$$

$$= \sum_{n=1}^{N} \pi_{ni} \left(-1 + \frac{r_{ni}}{\pi_{ni}} \right) x_n^T = \sum_{n=1}^{N} (r_{ni} - \pi_{ni}) x_n^T.$$

1.5. Iterative algorithm:

1. Randomly initialize φ and θ. Compute π_{nk} and r_{nk}.
2. Repeat until convergence ($\Delta L < \text{epsilon}$).

   E-step: update π_{nk} and r_{nk}

   M-step: update θ and φ using gradients obtained in 1.4. (θ and φ could be updated with gradient descent algorithm).

   Compute L. If $\Delta L < \text{epsilon}$ => Algorithm converged.

II Part.

1.

Likelihood: $p(\tilde{y} \mid \tilde{X}, \varphi, \theta) = \prod_{n=1}^{N} p(y_n \mid x_n, \theta, \varphi) \cdot \prod_{m=1}^{M} p(y_m \mid x_m, \varphi, \theta) =$

$$= \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_{nk} \, Exp(y_n \mid \lambda = \exp(\theta_k^T x_n)) \cdot \prod_{m=1}^{M} Exp(y_m \mid \lambda = \exp((\theta z_m)^T \cdot x_m)$$

Or we can also write it as:

$$\prod_{n=1}^{N} \sum_{k=1}^{K} \pi_{nk} \, Exp(y_n \mid \lambda = \exp(\theta_k^T x_n)) \prod_{m=1}^{M} \sum_{k=1}^{K} \pi_{mk}^{z_{mk}} \left(Exp(y_m \mid \lambda = \exp(\theta_k^T \cdot x_m))\right)^{z_{mk}}$$

log-Likelihood:

$$\log p(\tilde{y} \mid \tilde{X}, \varphi, \theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_{nk} \, p(y_n \mid x_n, \theta_k = \theta z_n) +$$

$$+ \sum_{m=1}^{M} \log p(y_m \mid x_m, \theta_m = \theta z_m) =$$

$$= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_{nk} \, Exp(y_n \mid \lambda = \exp(\theta_k^T x_n)) + \sum_{m=1}^{M} \log Exp(y_m \mid \lambda = \exp((\theta z_m)^T x_m)) =$$

$$= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_{nk} \, Exp(y_n \mid \lambda = \exp(\theta_k^T x_n)) -$$

$$- \sum_{m=1}^{M} (\theta z_m)^T x_m - \exp((\theta z_m)^T x_m) \, y_m .$$

2.

$$\frac{d}{d\varphi_i} \log p(\tilde{y} \mid \tilde{X}, \varphi, \theta) = \frac{d}{d\varphi_i} \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_{nk} \, Exp(y_n \mid \lambda = \exp(\theta_k^T x_n)) =$$

$$= \sum_{n=1}^{N} (r_{ni} - \pi_{ni}) \, x_n^T .$$

$$\frac{d}{d\theta_i}\log p(\tilde{y}\mid\tilde{X},\varphi,\theta)=\sum_{n=1}^{N} r_{ni}\cdot\left(1-\exp(\theta_i^T x_n)y_n\right)x_n^T+$$
$$+\sum_{m=1}^{M}\mathbb{1}(z_{mi}=1)\cdot\left(1-y_m\exp\left((\theta z_m)^T x_m\right)\cdot x_m^T\right.$$

3. The derivatives $\frac{d}{d\varphi_i}$ do not change. The derivative $\frac{d}{d\theta_i}$ is using new data points. Though the individual expert is a linear model — the overall model is non-linear (or soft piece-wise linear).

№2

$\{x_i\}_{i=1}^N$, $x_i\in\mathbb{R}^D$ $\quad Ex_i=\vec{0}$, $S=U\Lambda U^T$

2.1.

a) $z_{ni}=\left[z_n=U_K^T\tilde{x}_n=U_K^T x_n\right]=u_i^T x_n$,

with u_i – i-th column of U, U_K consists of first K-columns of U.

b) Empirical Mean

$Ez_{ni}=Eu_i^T x_n=u_i^T Ex_n=\vec{0}$

c) Empirical Variance.

Let's find covariance C_{ij} and $Vариance(z_i)=C_{ii}$.

$$C_{ij}=\frac{1}{N}\sum_{n=1}^{N} z_n^{(i)}z_n^{(j)}=\frac{1}{N}\sum_{n=1}^{N} u_i^T x_n x_n^T u_i=\left[\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T=S\right]=$$

$= u_i^T S u_j$

So $\text{Variance}(z_i) = u_i^T S u_i$

d) $\text{Variance}(z_i) = u_i^T S u_i = u_i^T U \Lambda U^T u_i =$

$= \left[U^T U = U U^T = I \Rightarrow \right.$

$\left. \Rightarrow U_K^T U \Lambda U^T U_K = \Lambda_K \in \mathbb{R}^{K \times K} \text{ with diagonal elements of } \Lambda \right] = \lambda_i.$

U_K: K columns, D rows, $\left\{ \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \end{pmatrix} \right\}_K^D$

U_K^T: D columns, K rows, $\left\{ \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \end{pmatrix} \right\}_D^K$

e) The proportion of explained variance is equal to: $\dfrac{\sum_{i=1}^{K} \lambda_i}{\sum_{j=1}^{D} \lambda_j}$.

So K should be the smallest possible such that it satisfies:

$\dfrac{\sum_{i=1}^{K} \lambda_i}{\sum_{j=1}^{D} \lambda_j} \geq 0{,}99.$

2.2.

$$C = \frac{1}{N} \sum_{n=1}^{N} z_n z_n^T = \frac{1}{N} \sum_{n=1}^{N} U^T x_n \cdot (U^T x_n)^T = \frac{1}{N} \sum_{n=1}^{N} U^T x_n x_n^T U =$$

$$= U^T \left(\frac{1}{N} \sum_{n=1}^{N} x_n x_n^T \right) \cdot U \; U^T S U = U^T U \Lambda U^T U = \left[U \text{ - orthogonal} \Rightarrow U^T = U^{-1} \right] = \Lambda$$

$\Rightarrow c_{ij} = 0$ for $i \neq j \Rightarrow$ dimensions are de-correlated

2.3. $E z_{ni} = 0$ and $\text{Variance } z_{ni} = \lambda_i$

Then for $\tilde{z}_{ni} = \sqrt{\frac{\gamma}{\lambda_i}} z_{ni} + m$:

$$E \tilde{z}_{ni} = \sqrt{\frac{\gamma}{\lambda_i}} E z_{ni} + E m = m$$

$$\text{Var } \tilde{z}_{ni} = \frac{\gamma}{\lambda_i} \text{ Var } z_{ni} = \gamma.$$