Aleksei Maliutin
1224****
a*****n@student.uva.nl

Homework Assignment 4
Machine Learning 1, 18/19

2018-12-16

1 Mixture of Experts

In class you discussed and were introduced to mixture models as a way to perform unsupervised learning tasks, e.g. clustering. Mixture models are not limited to only unsupervised learning and can be similarly used for supervised learning. In this homework we will discuss and explore Mixtures of Experts (MoEs), a model that softly partitions the input space and learns a supervised model for each area.

$$\begin{aligned} p(y_n|\mathbf{x}_n, z_n, \mathbf{\Theta}) &= p(y_n|\mathbf{x}_n, \boldsymbol{\theta}_k = \mathbf{\Theta}\mathbf{z}_n) \\ &= \text{Exponential}(y_n|\lambda = \exp(\boldsymbol{\theta}_k^T\mathbf{x}_n)) \\ &= e^{\boldsymbol{\theta}_k^T\mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_k^T\mathbf{x}_n}} \end{aligned}$$

The flexibility of MoEs stem from the fact that there is a “routing” mechanism which determines which of the K experts is appropriate for a specific datapoint $\mathbf{x}_n$. As in this case we have a discrete set of K experts, a simple linear routing mechanism is the following:

$$p(z_n = k|\mathbf{x}_n, \mathbf{\Phi}) = \pi_{nk} = \frac{e^{\boldsymbol{\phi}_k^T\mathbf{x}_n}}{\sum_j e^{\boldsymbol{\phi}_j^T\mathbf{x}_n}}$$

$\mathbf{\Theta}$ is a matrix in $\mathbb{R}^{D\times K}$ that contains the D-dimensional column vector of parameters for each expert.
$\mathbf{\Phi}$ is a matrix in $\mathbb{R}^{D\times K}$ that contains all of the parameters of the routing function.
With this information answer the following questions:

1. (1 point)

Write down the likelihood of the entire dataset, $p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi})$ and take its log under the i.i.d. assumption
ANSWER:

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \prod_{n=1}^{N}\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \mathbf{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi}) \\ &= \prod_{n=1}^{N}\sum_{k=1}^{K} \pi_{nk} \cdot e^{\boldsymbol{\theta}_k^T\mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_k^T\mathbf{x}_n}} \end{aligned}$$

Hence, taking the log

$$\begin{aligned} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \sum_{n=1}^{N}\ln\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \mathbf{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi}) = \\ &= \sum_{n=1}^{N}\ln\sum_{k=1}^{K} \pi_{nk} \cdot e^{\boldsymbol{\theta}_k^T\mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_k^T\mathbf{x}_n}} \end{aligned}$$

2. (1 point)

Write down the posterior probability r_{ni} of expert i producing the label y for datapoint n. We will also refer to this as the **responsibility** of expert i for datapoint n.
ANSWER:

$$\begin{aligned} r_{ni} = p(z_n = i|y_n, \mathbf{x}_n) &= \frac{p(z_n = i|\mathbf{x}_n, \mathbf{\Phi}) \cdot p(y_n|\mathbf{x}_n, z_n, \mathbf{\Theta})}{p(y_n|\mathbf{x}_n, \mathbf{\Theta}, \mathbf{\Phi})} = \\ &= \frac{p(z_n = i|\mathbf{x}_n, \mathbf{\Phi}) \cdot p(y_n|\mathbf{x}_n, z_n = i, \mathbf{\Theta})}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \mathbf{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})} \\ &= \frac{\pi_{ni} \cdot e^{\boldsymbol{\theta}_i^T\mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_i^T\mathbf{x}_n}}}{\sum_{k=1}^{K} \pi_{nk} \cdot e^{\boldsymbol{\theta}_k^T\mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_k^T\mathbf{x}_n}}} \end{aligned}$$

3. (4 points)

Take the derivative of the log-likelihood w.r.t. the parameters of each expert $\boldsymbol{\theta}_i$ and the parameters of the routing mechanism for each expert $\boldsymbol{\phi}_i$. Do not substitute expressions for the probabilities but rather provide your answer in terms $p(y_n|\mathbf{x}_n, z_n, \boldsymbol{\Theta})$ and $p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})$.

ANSWER:

The derivative of the log-likelihood w.r.t. $\boldsymbol{\theta}_i$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}_i} \ln p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) &= \frac{\partial}{\partial \boldsymbol{\theta}_i} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi}) = \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}_i} \ln \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi}) = \\
&= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \boldsymbol{\theta}_i} \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})} \\
&= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \boldsymbol{\theta}_i} p(y_n|\mathbf{x}_n, z_n = i, \boldsymbol{\Theta}) \cdot p(z_n = i|\mathbf{x}_n, \boldsymbol{\Phi})}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})} \\
&= \sum_{n=1}^{N} \underbrace{\frac{p(z_n = i|\mathbf{x}_n, \boldsymbol{\Phi}) \cdot p(y_n|\mathbf{x}_n, z_n = i, \boldsymbol{\Theta})}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})}}_{r_{ni}} \cdot \frac{\partial}{\partial \boldsymbol{\theta}_i} \ln p(y_n|\mathbf{x}_n, z_n = i, \boldsymbol{\Theta}) \\
&= \sum_{n=1}^{N} r_{ni} \cdot \frac{\partial}{\partial \boldsymbol{\theta}_i} \ln p(y_n|\mathbf{x}_n, z_n = i, \boldsymbol{\Theta})
\end{aligned}$$

The derivative of the log-likelihood w.r.t. $\boldsymbol{\phi}_i$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\phi}_i} \ln p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) &= \frac{\partial}{\partial \boldsymbol{\phi}_i} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi}) = \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{\phi}_i} \ln \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi}) = \\
&= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \boldsymbol{\phi}_i} \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})} \\
&= \sum_{n=1}^{N} \frac{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot \frac{\partial}{\partial \boldsymbol{\phi}_i} p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \underbrace{\frac{p(y_n|\mathbf{x}_n, z_n = k, \boldsymbol{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})}{\sum_{j=1}^{K} p(y_n|\mathbf{x}_n, z_n = j, \boldsymbol{\Theta}) \cdot p(z_n = j|\mathbf{x}_n, \boldsymbol{\Phi})}}_{r_{nk}} \cdot \frac{\partial}{\partial \boldsymbol{\phi}_i} \ln p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi}) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \cdot \frac{\partial}{\partial \boldsymbol{\phi}_i} \ln p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})
\end{aligned}$$

4. (4 points)

Replace the expressions for each of the respective probability distributions and compute the final derivatives for $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_i$.

ANSWER: First of all, calculate the derivative of $\ln p(y_n|\mathbf{x}_n, z_n, \boldsymbol{\Theta})$ w.r.t. the parameters of each expert $\boldsymbol{\theta}_i$ and the derivative of $\ln p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})$ w.r.t. the parameters of the routing mechanism for each expert $\boldsymbol{\phi}_i$.

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}_i}\ln p(y_n|\mathbf{x}_n, z_n = i, \boldsymbol{\Theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}_i} ln\Big(e^{\boldsymbol{\theta}_i^\mathsf{T}\mathbf{x}_n}\cdot e^{-y_n\cdot e^{\boldsymbol{\theta}_i^\mathsf{T}\mathbf{x}_n}}\Big) = \\ &= \frac{\partial}{\partial\boldsymbol{\theta}_i}\Big(\boldsymbol{\theta}_i^\mathsf{T}\mathbf{x}_n - y_n\cdot e^{\boldsymbol{\theta}_i^\mathsf{T}\mathbf{x}_n}\Big) \\ &= \mathbf{x}_n^\mathsf{T}(1 - y_n\cdot e^{\boldsymbol{\theta}_i^\mathsf{T}\mathbf{x}_n})\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\phi}_i}\ln p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi}) &= \frac{\partial}{\partial\boldsymbol{\phi}_i}\ln\Big(\frac{e^{\boldsymbol{\phi}_k^T\mathbf{x}_n}}{\sum_j e^{\boldsymbol{\phi}_j^T\mathbf{x}_n}}\Big) = \\ &= \frac{\partial}{\partial\boldsymbol{\phi}_i}\Big(\boldsymbol{\phi}_k^T\mathbf{x}_n - \ln\sum_j e^{\boldsymbol{\phi}_j^T\mathbf{x}_n}\Big) \\ &= \mathbf{I}_{ki}\mathbf{x}_n^\mathsf{T} - \frac{e^{\boldsymbol{\phi}_i^\mathsf{T}\mathbf{x}_n}}{\sum_j e^{\boldsymbol{\phi}_j^\mathsf{T}\mathbf{x}_n}}\cdot\mathbf{x}_n^\mathsf{T} \\ &= \mathbf{x}_n^\mathsf{T}\cdot\Big(\mathbf{I}_{ki} - \underbrace{\frac{e^{\boldsymbol{\phi}_i^T\mathbf{x}_n}}{\sum_j e^{\boldsymbol{\phi}_j^T\mathbf{x}_n}}}_{p(z_n=i|\mathbf{x}_n,\boldsymbol{\Phi})}\Big) = \\ &= \mathbf{x}_n^\mathsf{T}\cdot(\mathbf{I}_{ki} - p(z_n = i|\mathbf{x}_n, \boldsymbol{\Phi})) \\ &= \mathbf{x}_n^\mathsf{T}\cdot(\mathbf{I}_{ki} - \pi_{ni})\end{aligned}$$

Now, we can write the derivative of the log-likelihood w.r.t. $\boldsymbol{\theta}_i$ using formulas above:

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}_i}\ln p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) &= \sum_{n=1}^{N} r_{ni}\cdot\frac{\partial}{\partial\boldsymbol{\theta}_i}\ln p(y_n|\mathbf{x}_n, z_n = i, \boldsymbol{\Theta}) \\ &= \sum_{n=1}^{N} r_{ni}\cdot\mathbf{x}_n^\mathsf{T}(1 - y_n\cdot e^{\boldsymbol{\theta}_i^\mathsf{T}\mathbf{x}_n})\end{aligned}$$

and finally, the derivative of the log-likelihood w.r.t. $\boldsymbol{\phi}_i$:

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\phi}_i}\ln p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\cdot\frac{\partial}{\partial\boldsymbol{\phi}_i}\ln p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi}) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\cdot\mathbf{x}_n^\mathsf{T}\cdot(\mathbf{I}_{ki} - \pi_{ni}) \\ &= \sum_{n=1}^{N}\mathbf{x}_n^\mathsf{T}\Big(\underbrace{\sum_{k=1}^{K} r_{nk}\mathbf{I}_{ki}}_{=r_{ni}} - \pi_{ni}\cdot\underbrace{\sum_{k=1}^{K} r_{nk}}_{=1}\Big) \\ &= \sum_{n=1}^{N}\mathbf{x}_n^\mathsf{T}\cdot(r_{ni} - \pi_{ni})\end{aligned}$$

5. (1 point)

Write down an iterative algorithm that maximizes the log-probability of the data by jointly optimizing the $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ parameters. Make use of appropriate convergence criteria.
ANSWER:
Terminate conditions: T - max number of iteration or ϵ - converge ratio.
Hyper-parameters: η - learning rate

(1) Initialize $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$

(2) Initialize η

(3) Initialize $t = 0, \epsilon = +\infty$

(4) For *n* from 1 to N

a) **Expectation step** - update the posterior probability with current $\mathbf{\Theta}$ and $\mathbf{\Phi}$

   i. Calculate $\boldsymbol{\pi}_n = [\pi_{n1}, ..., \pi_{nK}]$

   ii. Calculate $\boldsymbol{r}_n = [r_{n1}, ..., r_{nK}]$

   iii. $t++$

b) **Maximization step** - update parameters $\mathbf{\Theta}$ and $\mathbf{\Phi}$ with SGD and derivatives from Question 4.

$$\mathbf{\Phi}^{(\tau+1)} = \mathbf{\Phi}^{(\tau)} + \eta \cdot \nabla^T_{\mathbf{\Phi}} L(\mathbf{\Phi}^{(\tau)})$$
$$\mathbf{\Theta}^{(\tau+1)} = \mathbf{\Theta}^{(\tau)} + \eta \cdot \nabla^T_{\mathbf{\Theta}} L(\mathbf{\Theta}^{(\tau)})$$

   NOTE: we use transpose here, because gradient in our notation is row-vector.

c) (Optional) Decrease η

d) (Optional) Earlier termination: if $\epsilon > \triangle L$ or $t > T$

(5) Repeat step 4 until $\epsilon > \triangle L$

(6) Return current $\mathbf{\Theta}$ and $\mathbf{\Phi}$

Assume that an oracle is available that provides you an extra set of **M** points that also has the information about which expert t should be employed.

1. (1 point)

Write down the likelihood of this extended dataset and take its log.
ANSWER:
NOTE: Let's mark "which expert t should be employed" for datapoint *n* as t_n

$$p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) = \underbrace{\prod_{n=1}^{N} \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \mathbf{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})}_{\text{original datapoints}} \cdot \underbrace{\prod_{n=N+1}^{N+M} p(y_n|\mathbf{x}_n, z_n = t_n, \mathbf{\Theta})}_{\text{extra points by oracle}} =$$

$$= \underbrace{\prod_{n=1}^{N} \sum_{k=1}^{K} \pi_{nk} \cdot e^{\boldsymbol{\theta}_k^T \mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_k^T \mathbf{x}_n}}}_{\text{original datapoints}} \cdot \underbrace{\prod_{n=N+1}^{N+M} e^{\boldsymbol{\theta}_{t_n}^T \mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_{t_n}^T \mathbf{x}_n}}}_{\text{extra points by oracle}}$$

Hence, taking the log

$$\ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) = \underbrace{\sum_{n=1}^{N} \ln \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, z_n = k, \mathbf{\Theta}) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})}_{\text{original datapoints}} + \underbrace{\sum_{n=N+1}^{N+M} \ln p(y_n|\mathbf{x}_n, z_n = t_n, \mathbf{\Theta})}_{\text{extra points by oracle}} =$$

$$= \underbrace{\sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_{nk} \cdot e^{\boldsymbol{\theta}_k^T \mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_k^T \mathbf{x}_n}}}_{\text{original datapoints}} + \underbrace{\sum_{n=N+1}^{N+M} \boldsymbol{\theta}_{t_n}^\mathsf{T} \mathbf{x}_n - y_n \cdot e^{\boldsymbol{\theta}_{t_n}^\mathsf{T} \mathbf{x}_n}}_{\text{extra points by oracle}}$$

2. (2 point)

Take the derivatives of this new log-likelihood w.r.t the parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_i$.
ANSWER:
As we see above, new log-likeligood is just a sum of old one and extra datapoints, hence, for the derivatives we can use formulas from previous step.
In case of derivative w.r.t. $\boldsymbol{\theta}_i$ - the parameters of each expert, we have changes only if $t_n = i$.

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) = \frac{\partial}{\partial \boldsymbol{\theta}_i} \left(\underbrace{\sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_{nk} \cdot e^{\boldsymbol{\theta}_k^T \mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_k^T \mathbf{x}_n}}}_{\text{original datapoints}} + \underbrace{\sum_{n=N+1}^{N+M} \boldsymbol{\theta}_{t_n}^\mathsf{T} \mathbf{x}_n - y_n \cdot e^{\boldsymbol{\theta}_{t_n}^\mathsf{T} \mathbf{x}_n}}_{\text{extra points by oracle}} \right)$$

$$= \sum_{n=1}^{N} r_{ni} \cdot \mathbf{x}_n^\mathsf{T} (1 - y_n \cdot e^{\boldsymbol{\theta}_i^\mathsf{T} \mathbf{x}_n}) \quad + \quad \sum_{n=N+1}^{N+M} \mathbf{I}_{t_n i} \cdot \mathbf{x}_n^\mathsf{T} (1 - y_n \cdot e^{\boldsymbol{\theta}_i^\mathsf{T} \mathbf{x}_n})$$

In case of derivative w.r.t. ϕ_i - the parameters of the routing mechanism for each expert, we don't have any changes, because extra datapoints are independent of them.

$$\frac{\partial}{\partial \boldsymbol{\phi}_i} \ln p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \frac{\partial}{\partial \boldsymbol{\phi}_i} \left(\underbrace{\sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_{nk} \cdot e^{\boldsymbol{\theta}_k^T \mathbf{x}_n} \cdot e^{-y_n \cdot e^{\boldsymbol{\theta}_k^T \mathbf{x}_n}}}_{\text{original datapoints}} + \underbrace{\sum_{n=N+1}^{N+M} \boldsymbol{\theta}_t^{\mathsf{T}} \mathbf{x}_n - y_n \cdot e^{\boldsymbol{\theta}_t^{\mathsf{T}} \mathbf{x}_n}}_{\text{extra points, do not depend on } \boldsymbol{\Phi}} \right)$$

$$= \sum_{n=1}^{N} \mathbf{x}_n^{\mathsf{T}} \cdot (r_{ni} - \pi_{ni})$$

3. (2 point)

What is the difference between the derivatives computed here and the derivatives computed previously? Is the overall model a linear or a non-linear one?
ANSWER:
As it was mentioned earlier, there are no significant changes in derivatives. However, the new ones are at least no less than old ones w.r.t. $\boldsymbol{\theta}_i$ and equal w.r.t. ϕ_i. It means, that adding *known* points to dataset leads to raster convergence. What is clear, if we know that expert *t* is the best for some point *n*, then for nearest unknown datapoint we are focusing also on optimization of that particular expert. In formal words, adding *known* points to dataset allows us to find optimal *expert covering* faster.
The overal model still be linear, because we use only linear terms in case of input variables - $\mathbf{x}_n$. For instance, each expert has a linear predictive model; moreover, calculation of $\boldsymbol{\pi}_n$ is also linear. Hence, the model does not depend on higher orders of $\mathbf{x}_n$ what makes it linear.

2 Principal Component Analysis

Suppose we have a data set $\{\mathbf{x}_1, .., \mathbf{x}_n\}$ of D-dimensional vectors, which have a zero mean for each dimension. Assume we perform a complete eigenvalue decomposition of the empirical covariance matrix $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$

1.

Initially, you are interested in only a single projection of your data such that the variance of this projection is maximized. Let $\mathbf{u}_i$ be the direction vector of a particular projection. Assume that $\mathbf{u}_i^T \mathbf{u}_i = 1$
ANSWER:

- (a) What is the projection z_{ni} of a given point $\mathbf{x}_n$ under the particular vector $\mathbf{u}_i$ (1 pt.)

$$z_{ni} = \mathbf{u}_i^{\mathsf{T}} \mathbf{x}_n$$

- (b) What is the empirical mean of the projection z_i across all points $\mathbf{x}_n$ (1 pt.)

According to the definition, the mean of projected data is a projection of the mean of the original dataset, but in our case we have "zero mean for each dimension", in other words, our dataset is *zero-centered*.

$$\bar{z}_i = \frac{1}{N} \sum_{n=0}^{N} z_{ni} = \mathbf{u}_i^{\mathsf{T}} \underbrace{\bar{\mathbf{x}}_n}_{=\mathbf{0}} = 0$$

- (c) What is the empirical variance of the projection z_i? Provide your answer in terms of the empirical covariance matrix $\mathbf{S}$ (1 pt.)

$$Var(\mathbf{z}_i) = \mathbf{u}_i^{\mathsf{T}} \mathbf{S} \mathbf{u}_i$$

- (d) Replace S with its eigenvalue decomposition and simplify the aforementioned expression. What is the variance now? (2 pt.)

$$Var(\mathbf{z}_i) = \mathbf{u}_i^{\mathsf{T}} \underbrace{\mathbf{S}}_{=\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}}} \mathbf{u}_i = \mathbf{u}_i^{\mathsf{T}} \mathbf{U}\boldsymbol{\Lambda} \underbrace{\mathbf{U}^T \mathbf{u}_i}_{\begin{bmatrix} 0_0 \\ .. \\ 1_i \\ .. \\ 0_D \end{bmatrix}} = \boldsymbol{\Lambda}_{ii} = \lambda_i = i\text{-th eigen value}$$

- (d) Suppose that you are interesting in reducing the dimensionality from D to K, such that 99% of the variance is maintained. How can you select an appropriate K? (2 pt.)

(1) Get the trace of covariance matrix $Tr(\mathbf{S})$ or in case of eigen-decomposition $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ simply $diag(\mathbf{\Lambda})$.

(2) Sort eigen-values that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D$

(3) Find an optimal K which is minimal for following condition:

$$\frac{\sum_{j=1}^{K} \lambda_j}{\underbrace{\sum_{j=1}^{D} \lambda_j}_{=Tr(\mathbf{S})}} - 0.99 \geq 0$$

2. (2 points)

Consider the projections of your data along the K principal components. Prove that the dimensions of the projections are de-correlated (Hint: check the value of the empirical covariance w.r.t. dimension i and j).
ANSWER:
Naive assumption: according to the lecture, the covariance matrix of the projected data is diagonal:

$$\frac{1}{N}\sum_{n=1}^{N} \mathbf{z}_n\mathbf{z}_n^{\mathsf{T}} = \mathbf{U}_K^{\mathsf{T}}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}\mathbf{U}_K = \mathbf{\Lambda}_M$$

Hence, from matrix $\mathbf{\Lambda}_M$ it is clear, that different dimensions of projected space are de-correlated due to the equality to zero in $\mathbf{\Lambda}_M$.
We know that projection of datapoint $\mathbf{x}_n$ along with the K principal components is

$$\mathbf{z}_n = \mathbf{U}_k^{\mathsf{T}}\mathbf{x}_n$$

hence, the projections on i and j coordinate of projected space respectively are

$$\mathbf{z}_{n_i} = \mathbf{u}_i^{\mathsf{T}}\mathbf{x}_n$$
$$\mathbf{z}_{n_j} = \mathbf{u}_j^{\mathsf{T}}\mathbf{x}_n$$

and both of them are scalars. Now, we can calculate the covariance of between dataset projected on on i coordinate and dataset projected j coordinate.

$$\text{Cov}(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{N}\sum_{n=1}^{N}(z_{n_i} - E(\mathbf{z}_i)) \cdot (z_{n_j} - E(\mathbf{z}_j)) = \frac{1}{N}\sum_{n=1}^{N}(z_{n_i} - 0) \cdot (z_{n_j} - 0) = \frac{1}{N}\sum_{n=1}^{N} z_{n_i} \cdot z_{n_j}$$
$$= \frac{1}{N}\sum_{n=1}^{N} \underbrace{\mathbf{u}_i^{\mathsf{T}}\mathbf{x}_n}_{\text{scalar}}\underbrace{\mathbf{x}_n^{\mathsf{T}}\mathbf{u}_j}_{\text{scalar}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{u}_i^{\mathsf{T}}\mathbf{u}_j\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n = \underbrace{\mathbf{u}_i^{\mathsf{T}}\mathbf{u}_j}_{\mathbb{I}(i=j)} \cdot \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n$$

As we see, the formula above is equal to zero if $i \neq j$. **Q.E.D.**

3. (3 points)

Imagine that you want to de-correlate all of the dimensions but still want to enforce a mean of $\mathbf{m}$ and variance of τ across the D dimensions. How can you post-process your projections such that they satisfy these properties?
ANSWER:
One of the side effect of PCA is de-correlation, hence, simply using PCA with *D* (the current dimensional size) principal component, we will get de-correlated data in new projected space.

$$\mathbf{z}_n = \mathbf{U}_D^{\mathsf{T}}\mathbf{x}_n$$

Then, to be sure that our dataset has a variance of τ across the D dimensions, first of all, we should rescale it to unit standard deviation

$$\mathbf{z}_n = \mathbf{\Lambda}_D^{-\frac{1}{2}}\mathbf{U}_D^{\mathsf{T}}\mathbf{x}_n$$

and then, upscale with $\sqrt{\tau}$ factor.

$$\mathbf{z}_n = \underbrace{\sqrt{\tau} \cdot \mathbf{I}_D}_{\text{update variance}} \left(\mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{U}_D^\mathsf{T} \mathbf{x}_n\right)$$

Last, to enforce a mean of $\mathbf{m}$ across the D dimensions, we can simply add it to each datapoint, as a result we'll get:

$$\mathbf{z}_n = \sqrt{\tau} \cdot \mathbf{I}_D \left(\mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{U}_D^\mathsf{T} \mathbf{x}_n\right) + \mathbf{m}$$

Show that these hold by computing the empirical mean and variance across the dataset.

$$\begin{aligned}\bar{\mathbf{z}}_n = \frac{1}{N}\sum_{n=1}^{N} \mathbf{z}_n &= \frac{1}{N}\sum_{n=1}^{N} \sqrt{\tau} \cdot \mathbf{I}_D \left(\mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{U}_D^\mathsf{T} \mathbf{x}_n\right) + \mathbf{m} \\ &= \underbrace{\frac{1}{N}\sum_{n=1}^{N} \sqrt{\tau} \cdot \mathbf{I}_D \left(\mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{U}_D^\mathsf{T} \mathbf{x}_n\right)}_{=\mathbf{0}} + \frac{1}{N}\sum_{n=1}^{N} \mathbf{m} = \mathbf{m}\end{aligned}$$

Q.E.D.
And for variance:

$$\begin{aligned}\mathrm{Var}(\mathbf{z}_n) = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{z}_n - \mathbf{m})^2 &= \frac{1}{N}\sum_{n=1}^{N} \left(\sqrt{\tau} \cdot \mathbf{I}_D \mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{U}_D^\mathsf{T} \mathbf{x}_n\right)^2 \\ &= \frac{1}{N}\sum_{n=1}^{N} \tau \left(\cdot \mathbf{I}_D \mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{U}_D^\mathsf{T} \mathbf{x}_n\right)^2 \\ &= \frac{1}{N}\sum_{n=1}^{N} \tau \cdot \mathbf{I}_D \mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{U}_D^\mathsf{T} \mathbf{x}_n \mathbf{x}_n^\mathsf{T} \mathbf{U}_D \mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{I}_D \\ &= \underbrace{\tau \cdot \mathbf{I}_D \mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{U}_D^\mathsf{T} \left(\frac{1}{N}\sum_{n=1}^{N} \cdot \mathbf{x}_n \mathbf{x}_n^\mathsf{T}\right) \mathbf{U}_D \mathbf{\Lambda}_D^{-\frac{1}{2}} \mathbf{I}_D}_{=\mathbf{I}_D} \\ &= \tau \cdot \mathbf{I}_D\end{aligned}$$

Q.E.D.