— Solution notes —

Fifth practice exercises in Machine learning 1 – 2025 – Paper 1

1 Principal component analysis (October)

Suppose we have a data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of D-dimensional vectors, which have a zero mean for each dimension. Assume we perform a complete eigenvalue decomposition of the empirical covariance matrix $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. 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 $\mathbf{u}_i^T\mathbf{u}_i = 1$.

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

Answer:

The projection of the vector $\mathbf{x}_n$ over the vector $\mathbf{u_i}$ is given by:

$$z_{ni} = \mathbf{u_i}^T\mathbf{x}_n$$

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

Answer: The empirical mean:

$$E[z_i] = \frac{1}{N}\sum_{n=1}^{N}\mathbf{u}_i^T\mathbf{x}_n = \mathbf{u}_i^T\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\right) = 0$$

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

Answer:

The variance of a variable $\mathbf{z}$ is defined as:

$$V[z_i] = E[(z_i - \bar{z}_i)^2]$$

Hence, in our case:

$$V[z_i] = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_i^T\mathbf{x}_n)(\mathbf{u}_i^T\mathbf{x}_n)^T$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbf{u}_i^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{u}_i$$

$$= \mathbf{u}_i^T\mathbf{S}\mathbf{u}_i$$

(d) Replace $\mathbf{S}$ with its eigenvalue decomposition and simplify the aforementioned expression. What is the variance now?

Answer:

$$V[z_i] = \mathbf{u}_i^T \mathbf{S} \mathbf{u}_i = \mathbf{u}_i^T \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \mathbf{u}_i$$

Since $\mathbf{u}_i^T \mathbf{u}_i = 1$ and $\mathbf{u}_i^T \mathbf{u}_j = 0$ then we can re-write:

$$= \mathbf{e}_i^T \Lambda \mathbf{e}_i = \lambda_i$$

with $\mathbf{e}_i$ to be a vector with zeros except the position with index i.

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

Answer: We need to sort eigenvalues in descending order. Then by picking K largest eigenvalues using the following formula:

$$\frac{\sum_{i=1}^{K-1} \lambda_i}{\sum_{i=1}^{D} \lambda_i} < 0.99 \leq \frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{D} \lambda_i}$$

Fifth practice exercises in Machine learning 1 – 2025 – Paper 1

2 Introduction to backpropagation (October)

Consider a two-layer neural network, defined as follows

$$\mathbf{z}^{(1)} = \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}, \quad \mathbf{a}^{(1)} = \tanh(\mathbf{z}^{(1)}),$$
$$\mathbf{z}^{(2)} = \mathbf{W}^{(2)}\mathbf{a}^{(1)} + \mathbf{b}^{(2)}, \quad \hat{\mathbf{y}} = \text{softmax}(\mathbf{z}^{(2)}),$$

where

- $\mathbf{x} \in \mathbb{R}^D$ is a single input example.
- $\hat{\mathbf{y}} \in \mathbb{R}^K$ is the predicted output vector.
- The first hidden layer (1) has H neurons.

(*a*) What are the shapes of $\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{a}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}$?

Answer:

$$\mathbf{W}^{(1)} \in \mathbb{R}^{H\times D}, \quad \mathbf{b}^{(1)} \in \mathbb{R}^{H\times 1}, \quad \mathbf{a}^{(1)} \in \mathbb{R}^{H\times 1}, \quad \mathbf{W}^{(2)} \in \mathbb{R}^{K\times H}, \quad \mathbf{b}^{(2)} \in \mathbb{R}^{K\times 1}.$$

(*b*) We feed the neural network with a sample $\mathbf{x}$, obtaining the prediction $\hat{\mathbf{y}}$. We assume we have the ground truth label $\mathbf{y}$ for that sample $\mathbf{x}$. Write down the formula for the cross-entropy loss $L(\mathbf{y}, \hat{\mathbf{y}})$ for that sample.

Answer:

$$L(\mathbf{x}, \mathbf{y}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k.$$

(*c*) Compute $\frac{\partial L}{\partial \hat{y}_i}$.

Answer: The loss for one example is

$$L = -\sum_{j=1}^{K} y_j \log \hat{y}_j,$$

so

$$\frac{\partial L}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}.$$

(*d*) Compute $\frac{\partial L}{\partial z_i{}^{(2)}}$.

Answer: By the chain rule, we have

$$\frac{\partial L}{\partial z_i^{(2)}} = \frac{\partial L}{\partial \hat{\mathbf{y}}} \frac{\partial \mathbf{y}}{\partial z_i^{(2)}} = \sum_{j=1}^{K} \frac{\partial L}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_i^{(2)}}.$$

From the previous question, we know that

$$\frac{\partial L}{\partial \hat{y}_j} = -\frac{y_j}{\hat{y}_j}.$$

The derivative of the softmax (computed in HW1) is

$$\frac{\partial \hat{y}_j}{\partial z_i^{(2)}} = \hat{y}_j(\delta_{ij} - \hat{y}_i)$$

Plugging everything in, we obtain

$$\frac{\partial L}{\partial z_i^{(2)}} = \sum_{j=1}^{K} \left(-\frac{y_j}{\hat{y}_j}\right) \hat{y}_j(\delta_{ij} - \hat{y}_i) = -\sum_{j=1}^{K} y_j \delta_{ij} + \hat{y}_i \sum_{j=1}^{K} y_j = \hat{y}_i - y_i.$$

(*e*) Compute $\frac{\partial L}{\partial W_{ij}^{(2)}}$.

Answer:

By the chain rule,

$$\frac{\partial L}{\partial W_{ij}^{(2)}} = \frac{\partial L}{\partial z_i^{(2)}} \cdot \frac{\partial z_i^{(2)}}{\partial W_{ij}^{(2)}}.$$

From the previous question, we know

$$\frac{\partial L}{\partial z_i^{(2)}} = \hat{y}_i - y_i.$$

Recall that

$$z_i^{(2)} = \sum_{j=1}^{H} W_{ij}^{(2)} a_j^{(1)} + b_i^{(2)},$$

and therefore

$$\frac{\partial z_i^{(2)}}{\partial W_{ij}^{(2)}} = a_j^{(1)}.$$

Finally,

$$\boxed{\frac{\partial L}{\partial W_{ij}^{(2)}} = (\hat{y}_i - y_i)\, a_j^{(1)}.}$$
