Intro to Machine Learning Homework Assignment 6

Goktug Saatcioglu

NetId: gs2417

1. (a) A common phenomenon with high-dimensional data vectors is that many machine learning algorithms may perform poorly due to the curse of dimensionality. The curse of dimensionality referes to the phenomenon of the generalization error first decreasing until a point as the dimensions of the data are increased and then starting to increase as even more dimensions are added. This happens because as we increase the dimensions of our data the data becomes more sparse which means we need more data observations for training in order to avoid overfitting for our model. However, most of the time increasing the number of observations is not a viable strategy as we would need huge amount of data points to process high dimensional data without overfitting. Thus, reducing the data to a lower dimensions allows us to avoid the curse of dimensionality and is a reason why it is more efficient to process data points if they are lower dimensional vectors.

(b) A potential trouble of reducing the dimensionality of the input vectors before training a classifier is that the reduced features are not easily interpretable for many of the algorithms we use for dimensionality reduction. This is especially true for both PCA and SVD and training on the reduced dataset makes it harder to interpret the results. A solution could be using a different algorithm for reduction where the reduced features are more interpretable. But then there still will be the issue of what constraints to choose including the question of how many features should we reduce the dimensions to. An aggressive approach could lead to a loss of crucial information which then could lead to a model that does not generalize well. We see that both the lack of interpretability of the reduced features and the choice of constraints including what dimensions to reduce to are potential troubles of reducing the dimensionality of the input vectors before training a classifier.

2. (a) We consider the reconstruction error over the whole training set D where N is the size of the set, d is the original dimension of the set and q is the dimension we reduce to (i.e. the q eigenvectors components chosen). We can reconstruct any data vector x^i from its reduced vector z^i as follows:

$$\hat{x}^i = \sum_{j=1}^{q} z_j^i w_j.$$

The reconstruction error is then defined as

$$\frac{1}{N}\sum_{i=1}^{N} \left\|x^i - \hat{x}^i\right\|_2^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}(x_j^i - \hat{x}_j^i)^2$$

and it easy to see that if $q = d$ then

$$\hat{x}^i = x^i = \sum_{j=1}^{d} z_j^i w_j$$

for all i which gets us zero-reconstruction error. We also note that if D is mean-centered such that $D = \{d^1, \ldots, d^N\}$ is tranformed to $D = \{d^1 - \bar{d^1}, \ldots, d^N - \bar{d^N}\} = \{x^1, \ldots, x^N\}$ (where $\bar{d^i}$ is

the mean of data vector d^i) then the covariance matrix C is given by

$$C = \frac{1}{N}\sum_{i=1}^{N} x^i (x^i)^\top. \tag{1}$$

Using the re-writing of x^i and (1) we get:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\left\|x^i - \hat{x}^i\right\|_2^2 &= \frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{j=1}^{d} z_j^i w_j - \sum_{j=1}^{q} z_j^i w_j\right\|_2^2 && \text{by definition}\\
&= \frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{j=1}^{q}(z_j^i - z_j^i)w_j + \sum_{j=q+1}^{d} z_j^i w_j\right\|_2^2 && \text{re-arrange sums}\\
&= \frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{j=q+1}^{d} z_j^i w_j\right\|_2^2 && \text{simplify}\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=q+1}^{d}(z_j^i)^2\left\|w_j\right\|^2 && \text{by definition}\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=q+1}^{d}(z_j^i)^2 && \text{since } \|w_j\|^2 = 1\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=q+1}^{d}(w_j^\top x^i)((x^i)^\top w_j) && \text{by definition}\\
&= \frac{1}{N}\sum_{j=q+1}^{d} w_j^\top (\sum_{i=1}^{N} x^i (x^i)^\top) w_j && \text{re-arrange sums}\\
&= \frac{1}{N}\sum_{j=q+1}^{d} w_j^\top N C w_j && \text{using (1)}\\
&= \sum_{j=q+1}^{d} w_j^\top C w_j && \text{take out the constant}\\
\therefore \frac{1}{N}\sum_{i=1}^{N}\left\|x^i - \hat{x}^i\right\|_2^2 &= \sum_{j=q+1}^{d} w_j^\top C w_j,
\end{aligned}$$

which proves the property we seek to prove.

(b) We again assume that the data has been mean centered and (1) holds. Given some data matrix $X = \{x_1, \ldots, x_d\}$ the covariance matrix C of X is defined as $C = XX^\top$ and the size of the dataset X is given by $|X| = d$. To prove the statement in the question we must show that it holds for both the forward direction ($\implies$) and the backward direction ($\impliedby$). Beginning with the forward direction ($\implies$), we know that to find the eigendecomposition of a square matrix for PCA the matrix must be diagonalizable. Thus, if W is the matrix that diagonalizes C then the eigenvalues of C can be obtained by

$$W^{-1} C W = \Sigma$$

where W is the matrix of the eigenvectors of C. By the construction of C we know that in this case $W^{-1} = W^T$ (meaning W is orthogonal) and this fact can be proven through the equivalence of the eigendecomposition of C and the singular value decomposition of C. Using this fact we get

$$W^\top C W = \Sigma$$

which is enough to prove the forward direction. However, for the sake of completeness consider the following derivation:

$$\begin{aligned}
W^\top C W &= W^\top X X^\top W && \text{since } C = XX^\top \\
&= Z X^\top W && \text{since } Z = W^\top X \\
&= Z Z^\top && \text{since } Z^\top = X^\top W \\
&= \sum_{j=1}^{d} z_j z_j^\top && \text{by definition} \\
&= \sum_{j=1}^{d} w_j^\top x_j x_j^\top w_j && \text{by definition} \\
&= \sum_{j=1}^{d} w_j^\top C w_j && \text{by definition of } C \\
\therefore W^\top C W &= \sum_{j=1}^{d} w_j^\top C w_j.
\end{aligned}$$

Since we know that $W^\top C W = \sum_{j=1}^{d} w_j^\top C w_j$ and that $W^\top C W = \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$ we conclude that

$$\Sigma = W^\top C W \implies \sigma_j^2 = w_j^\top C w_j, \text{ for all } j = 1, \ldots, d,$$

which finishes the $\implies$ portion of this proof. Now let's consider the backward direction ($\impliedby$), if

$$\sigma_j^2 = w_j^\top C w_j, \text{ for all } j = 1, \ldots, d$$

is true then the following derivation is possible:

$$\begin{aligned}
\Sigma &= \text{diag}(w_1^\top C w_1, \ldots, w_d^\top C w_d) && \text{by assumption} \\
&= \sum_{j=1}^{d} w_j^\top C w_j && \text{by definition} \\
&= \sum_{j=1}^{d} w_j^\top x_j x_j^\top w_j && \text{by definition of } C \\
&= \sum_{j=1}^{d} z_j z_j^\top && \text{by definition} \\
&= Z Z^\top && \text{by definition} \\
&= Z X^\top W && \text{since } Z^\top = X^\top W \\
&= W^\top X X^\top W && \text{since } Z = W^\top X \\
&= W^\top C W && \text{since } X X^\top = C \\
\therefore \Sigma &= W^\top C W.
\end{aligned}$$

This derivation completes $\impliedby$ portion of this proof. Since we have proved that both $\implies$ and $\impliedby$ holds we can say that the $\iff$ holds and we have shown the property we seek to show.

3. See hw6_pca.ipynb and hw6_nmf.ipynb.