PCA

Dimensional Reduction

- Usually considered an unsupervised learning method
- Used for learning the low-dimensional structures in the data (e.g., topic vectors instead of bag-of-words vectors, etc.)
- Fewer dimensions $\Rightarrow$ Less chances of overfitting $\Rightarrow$ Better generalization.

Linear Dimensionality Reduction

- Projection matrix $U = [u_1, u_2, \cdots, u_K]$ of size $D \times K$ defines K linear projection direction.
- U is to project $x^{(i)} \in \mathbb{R}^D$ to $z^{(i)} \in \mathbb{R}^K$

$$Z = U^T \cdot X, X = [x^{(1)} \cdots x^{(N)}] \in \mathbb{R}^{D \times N} Z = [z^{(1)} \cdots z^{(N)}] \in \mathbb{R}^{K \times N}$$

PCA

- Usage: s dimensionality reduction, lossy data compression, feature extraction, and data visualization

  Def. (2 commonly used definitions)

  - Learning projection directions that capture maximum variance in data
  - Learning projection directions that result in smallest reconstruction error

- Projection of $x^{(i)}$ along a one-dim subspace defined by $u_1 \in \mathbb{R}^D$, where $||u_1|| = 1$.
- Mean of projections is $u_1^T \mu$, where $\mu = \frac{1}{N} \sum_{i=1}^N x^{(i)}$ is the mean of all data.
- Variance of projections is $u_1^T S u_1$

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} \left(u_1^T x^{(i)} - u_1^T \mu\right)^2 &= \frac{1}{N} \sum_{i=1}^{N} \left[u_1^T \left(x^{(i)} - \mu\right)\right]^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \left[u_1^T \left(x^{(i)} - \mu\right)\right] \left[u_1^T \left(x^{(i)} - \mu\right)\right]^T \\
&= E\left[u_1^T \left(X - \mu\right) \left(X - \mu\right)^T u_1\right] \\
&= u_1^T S u_1
\end{aligned}$$

- S is the $D \times D$ data covariance matrix

$$S = E\left[(X - \mu)(X - \mu)^T\right] = \frac{1}{N} \sum_{i=1}^{N} \left(x^{(i)} - \mu\right) \left(x^{(i)} - \mu\right)^T$$

Optimization

- We want u_1 s.t. the variance of the projected data is maximized

$$\begin{aligned} \max_{u_1} \quad & u_1^T S u_1 \\ \text{s.t.} \quad & u_1^T u_1 = 1 \end{aligned}$$

- The method of Lagrange multipliers

$$\mathcal{L}\left(u_1, \lambda_1\right) = u_1^T S u_1 - \lambda_1 \left(u_1^T u_1 - 1\right)$$

- where λ_1 is a Lagrange multiplier - Take the derivative w.r.t. u_1 and setting to zero

$$\frac{\partial}{\partial u_1} \mathcal{L}\left(u_1, \lambda_1\right) = (S + S^T) u_1 - 2\lambda_1 u_1 = 0 \Leftrightarrow S u_1 = \lambda_1 u_1,\ (S = S^T)$$

- Thus, u_1 is an eigenvector of S
- The variance of projection is $u_1^T S u_1 = \lambda_1$.
- Variance is maximized when u_1 is the top eigenvector with largest eigenvalue (so-called the first Principle Component, PC).

Steps

1. Center the data (subtract μ for each data)
2. Compute the covariance matrix $S = \frac{1}{N} X X^T$
3. Perform eigen decomposition of S and take first K leading eigenvectors $\{u_i\}_{i=1,\cdots,K}$.
4. The projection is therefore given by $Z = U^T X$