12 Continuous Latent Variables

Definition. ***Motivation***

1. ***Idea*** *datasets have property that data points all lie close to a manifold of a much lower dimension than that of original data space*

2. ***Digit Example*** *translation, rotation, and scaling are latent variables. Additional degree of freedom of variability comes from variability in individual writing style*

3. ***PCA*** *A continuous latent model that assumes Gaussian distribution for both latent and observed variables and make use of linear-Gaussian dependence of observed variables on the state of the latent variables*

12.1 Principal Component Analysis

Definition. ***PCA*** *has 2 formulation*

1. *Orthogonal projection of data onto a lower dimensional linear space, the principal subspace, such that variance of projected data is maximized*

2. *Linear projection that minimizes the average projection cost, defined as the mean squared distance between the data points and their projections*

Definition. ***Maximum variance formulation*** *GIven* $\{\mathbf{x}_n\}$ *where* $\mathbf{x}_n \in \mathbb{R}^D$*. Goal is to project data onto space with dimensionality* $M < D$ *while maximizing varaince of projected data. Given* $M = 1$*, let* $\mathbf{u}$ *be a unit vector (*$\mathbf{u}^T\mathbf{u} = 1$*). Each data point is projected onto a scalar* $\mathbf{u}^T\mathbf{x}_n$*. Let mean of projected data be*

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

We want to maximize projected variance

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n^T\mathbf{u} - \overline{\mathbf{x}}^T\mathbf{u}\right)^2 &= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n^T\mathbf{u} - \overline{\mathbf{x}}^T\mathbf{u}\right)^T\left(\mathbf{x}_n^T\mathbf{u} - \overline{\mathbf{x}}^T\mathbf{u}\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{u}^T\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)^T\mathbf{u} \\ &= u^TSu\end{aligned}$$

where $S = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)^T$ *is the covariance matrix. Maximize using langrange multipliers*

$$\mathbf{u}^T\mathbf{S}\mathbf{u} + \lambda(1 - \mathbf{u}^T\mathbf{u})$$

gives us that variance will be maximized when we set u be eigenvector having largest eigenvalue λ. In general, the optimal linear projection for which the variance of projectedf data is maximized is defined by M eigenvectors $\mathbf{u}_1, \cdots, \mathbf{u}_M$ of the data covariance matrix $\mathbf{S}$ corresponding to M largest eigenvalues $\lambda_1, \cdots, \lambda_M$

Definition. ***Minimum-error formulation*** *Given orthonormal basis for data space $\{\mathbf{u}_1, \cdots, \mathbf{u}_M, \cdots, \mathbf{u}_D\}$, where the first M basis forms the basis for the principal subspace where we project onto. We approximate the each data point $\mathbf{x}_n$*

$$\mathbf{x}_n = \sum_{i=1}^{D} (\mathbf{x}_n^T \mathbf{u}_i)\mathbf{u}_i \qquad \xleftarrow{approximate} \qquad \tilde{\mathbf{x}}_n = \sum_{i=1}^{M} \alpha_{ni}\mathbf{u}_i + \sum_{i=M+1}^{D} \beta_i \mathbf{u}_i$$

*where α_{ni} varies and β_i fixed constant. Goal is to **minimize squared distance** between original data point $\mathbf{x}_n$ and its approximation $\tilde{\mathbf{x}}_n$, averaged over the entire dataset,*

$$J = \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2$$

computing $\frac{\partial J}{\partial \alpha_{ni}}$ and $\frac{\partial J}{\partial \beta_i}$, set to zero, we get

$$z_{ni} = \mathbf{x}_n^T \mathbf{u}_i \quad b_i = \overline{\mathbf{x}}^T \mathbf{u}_i \qquad \xrightarrow{reformulateJ} \qquad J = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} (\mathbf{x}_n^T \mathbf{u}_i - \overline{\mathbf{x}}^T \mathbf{u}_i)^2 = \sum_{i=M+1}^{D} \mathbf{u}_i^T \mathbf{S} \mathbf{u}_i$$

Similar to how we maximized $\mathbf{u}^T \mathbf{S} \mathbf{u}$ in the maximum variance formulation, the choice of choosing $\mathbf{u}_i$ where $i = M+1, \cdots, D$ where the corresponding eigenvalues are smallest minimizes J. Therefore the distortion measure can be written as

$$J = \sum_{i=M+1}^{D} \langle Su_i, u_i \rangle = \sum_{i=M+1}^{D} \lambda_i \langle u_i, u_i \rangle = \sum_{i=M+1}^{D} \lambda_i$$

Note this is equivalent to picking eigenvectors as basis for the principal component whose corresponding eigenvalues are the largest in the previous formulation

Definition. ***PCA for high-dimensional data***

1. ***Compute eigenvalues*** *Let $\mathbf{X}$ be $N \times D$ centered data matrix, whose n-th row given by $(\mathbf{x} - \overline{\mathbf{x}})^T$. The covariance is therefore $\mathbf{S} = N^{-1}\mathbf{X}^T\mathbf{X}$, then*

   $$\frac{1}{N}\mathbf{X}^T\mathbf{X}\mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad \xrightarrow{\times \mathbf{X}} \qquad \frac{1}{N}\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{u}_i) = \lambda_i(\mathbf{X}\mathbf{u}_i) \iff \frac{1}{N}\mathbf{X}\mathbf{X}^T\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

   for $\mathbf{v}_i = \mathbf{X}\mathbf{u}_i$. We can solve for the eigenvalues λ_i in $O(N^3)$ time instead of $O(D^3)$.

2. ***Compute eigenvectors*** *In order to determine the eigenvectors we multiply both sides by $\mathbf{X}^T$*

   $$(\frac{1}{N}\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{v}_i) = \lambda_i(\mathbf{X}^T\mathbf{v}_i)$$

   where $\mathbf{X}^T\mathbf{v}_i = \mathbf{u}_i$ an eigenvector of $\mathbf{S}$

12.4 Nonlinear Latent Variable Models

Definition. ***Independent component analysis*** *Observed variables related linearly to the latent variables, but for which the latent distribution is non-Gaussian*

Definition. ***Autoassociative neural networks (Autoencoders)***

1. ***Idea*** *A multiplayer perception where the input/output dimensions are equal D and that the hiddern dimension is smaller $M < D$ (act as a bottleneck layer). We want to minimize degree of mismatch between input vectors and their reconstruction, i.e.*

$$\mathcal{E}(\mathbf{w}) = \frac{1}{2} \sum_n \|y(\mathbf{x}_n, \mathbf{w}) - \mathbf{x}_n\|^2$$

2. ***Linear Dimensionality Reduction*** *autoassociative neural net with 1 hiddern layer performs projection onto M-dimensional subspace spanned by the first M principal components of the data. THe vector of weights leading into the hidden unit forms a basis set that spans the principal subspace*

3. ***Nonlinear Dimensionality Reduction*** *autoassociative neural net with more than 1 hidden unit containing nonlinear activation function does nonlinear dimensionality reduction.*