Session 5 reading of PRML: Ch. 12 Continuous Latent Variables 2/21/2022

assigned sections

- 12.0 – 12.2.3 except 12.2.2 & 12.2.3 is optional
- 12.2.4

12.1 Principal Component Analysis

12.1.1 Maximum variance formulation

$\{x_n\}$: data set of observations w/ $n = 1, \ldots, N$

x_n is D-dimensional

goal: project the data onto a space of dimensionality $M < D$ while maximizing the variance of the projected data.

consider projection to 1 dimension: $M = 1$

- define direction using D-dim vector u_1 that is a unit vector s.t. $u_1^T u_1 = 1$
- project each data point onto u_1: $u_1^T x_n$

mean of projected data = $\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$

var. of projected data = $\frac{1}{N} \sum_{n=1}^{N} \{u_1^T x_n - u_1^T \bar{x}\}^2 = u_1^T S u_1$

where $S = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$ ← (covariance matrix)

- maximize var. w.r.t. u_1

  - the unit vector constrains to solution & use Lagrange mult. method

$u_1^T S u_1 + \lambda_1 \underbrace{(1 - u_1^T u_1)}_{= 0}$

$\frac{d}{du}$ & set equal to 0: $S u_1 = \lambda_1 u_1$

$u_1^T S u_1 = \lambda_1$

$\{ \therefore$ maximize var. when u_1 = eigenvector of largest eigenvalue $\lambda_1 \}$

↳ "first principal component"

12.1.2. Minimum-error formulation

goal: minimize projection error

- introduce a complete set of D-dimensional orthonormal basis vectors $\{u_i\}$

$$u_i^T u_j = \delta_{ij} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases} \qquad \text{for } i = 1, \dots, D$$

- each data point in $\{x_n\}$ can be represented as a linear comb. of the basis vectors:

$$x_n = \sum_{i=1}^{D} \alpha_{ni} u_i \qquad \text{where } \alpha_{ni} \text{ are coef. for each data point}$$

- solve for α_{nj} → $\alpha_{nj} = x_n^T u_j$ ⇒ $x_n = \sum_{i=1}^{D} (x_n^T u_i) u_i$
- want to approx. this data point w/ $M < D$ variables
  - use the first M of the basis vectors:

$$\tilde{x}_n = \underbrace{\sum_{i=1}^{M} z_{ni} u_i}_{\text{vary per data point}} + \underbrace{\sum_{i=M+1}^{D} b_i u_i}_{\text{constant for all data points}}$$

- find $\{u_i\}, \{z_{ni}\}, \{b_i\}$ to minimize reconstruction error
  → we'll use squared dist. as the error:

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

- solve, solve, solve, ...

$$J = \sum_{i=M+1}^{D} u_i^T S u_i$$

- w/ the constraint of orthonormality on $\{u_i\}$, solution is expressed as the eigenvectors/values of the covariance matrix

12.2. Probabilistic PCA

- PCA can be expressed as the MLE solution of a probabilistic latent variable model

- Introduce a latent variable $\mathbf{z}$
  - corresponds to the principal-component subspace
  - normal prior: $p(\mathbf{z}) \sim N(\mathbf{z} \mid 0, \mathbf{I})$ "zero-mean, unit-cov. Gaussian"

- conditional dist of the observed variable $\mathbf{x}$: $p(\mathbf{x} \mid \mathbf{z}) = N(\mathbf{x} \mid \overbrace{\mathbf{W}}^{D \times M}\mathbf{z} + \overbrace{\mu}^{D}, \sigma^2 \mathbf{I})$
  - $\mathbf{x}$ is a general linear fxn of $\mathbf{z}$ governed by $\mathbf{W}_{D \times M}$ & D-dim μ

- generate a D-dim. observed variable $\mathbf{x}$: $\mathbf{x} = \mathbf{W}\mathbf{z} + \mu + \epsilon$

- estimate $\mathbf{W}, \mu, \sigma^2$ w/ MLE
  - need marginal dist. $p(\mathbf{x})$: $p(\mathbf{x}) = \int p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) \, d\mathbf{z}$
    $= N(\mathbf{x} \mid \mu, \mathbf{C})$
    where $\mathbf{C}$ is a $D \times D$ cov. matrix: $\underline{\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}}$

- redundancy in the current parameterization w/ rotations of the latent space coordinates

* on its own, could rotate W (causing the non-identif.) but WW^T is invariant to rotation $\Rightarrow$ C is invariant to rotation & has a unique value

Session 5 class notes 2/23/2022

Lecture notes : Dimensionality Reduction.

Probabilistic PCA

· build a generative model instead of defining an error fxn (regular PCA)

· For MLE solution, $\hat{\sigma}^2$ ends up as the average of eigenvalues of cov. mat. that are not w/in M

- intuitive b/c is basically just the variance not captured by PCA solution (ie. remaining dimensions)

Factor Analysis

· diff w/ PPCA is non-isotropic noise in the generative model of the data given the latent space

· when to use FA over PPCA :

1. depends on which generative model matches your problem
   - can make isotropic variance assumption?
2. PPCA is faster to fit
3. separate roles of W & Ψ in FA so get more info.