Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2011
Prof. Erik Sudderth

Lecture 21: EM for Factor Analysis

Many figures courtesy Kevin Murphy’s textbook,
Machine Learning: A Probabilistic Perspective
Principal Components Analysis (PCA)

3D Data

Best 2D Projection

Best 1D Projection
PCA as Low Rank Approximation

\[ \mathbf{X} \overset{D}{=} \mathbf{LZ} \overset{L}{=} \mathbf{W}' \]
Maximizes Variance & Minimizes Error

C. Bishop, *Pattern Recognition & Machine Learning*
Probabilistic PCA & Factor Analysis

- **Both Models**: Data is a linear function of low-dimensional latent coordinates, plus Gaussian noise

\[ p(x_i | z_i, \theta) = \mathcal{N}(x_i | Wz_i + \mu, \Psi) \quad p(z_i | \theta) = \mathcal{N}(z_i | 0, I) \]

- **Factor analysis**: \( \Psi \) is a general diagonal matrix
- **Probabilistic PCA**: \( \Psi \) is a multiple of identity matrix

C. Bishop, *Pattern Recognition & Machine Learning*
Lower Bounds on Marginal Likelihood

\[ \text{KL}(q||p) \]

\[ \mathcal{L}(q, \theta) \]

\[ \ln p(X|\theta) \]

C. Bishop, Pattern Recognition & Machine Learning
Expectation Maximization Algorithm

**E Step:** Optimize distribution on hidden variables given parameters

\[ \text{KL}(q||p) = 0 \]

\[ \mathcal{L}(q, \theta^{\text{old}}) \]

\[ \ln p(X|\theta^{\text{old}}) \]

\[ \mathcal{L}(q, \theta^{\text{new}}) \]

\[ \ln p(X|\theta^{\text{new}}) \]

**M Step:** Optimize parameters given distribution on hidden variables

C. Bishop, Pattern Recognition & Machine Learning
EM Algorithm for Probabilistic PCA

C. Bishop, Pattern Recognition & Machine Learning
EM Algorithm for Probabilistic PCA

E-Step #1 (Projection)
EM Algorithm for Probabilistic PCA

M-Step #1 (Regression)

C. Bishop, Pattern Recognition & Machine Learning
EM Algorithm for Probabilistic PCA

C. Bishop, Pattern Recognition & Machine Learning

E-Step #2 (Projection)
EM Algorithm for Probabilistic PCA

M-Step #2 (Regression)

C. Bishop, Pattern Recognition & Machine Learning
EM Algorithm for Probabilistic PCA

Converged Solution

C. Bishop, Pattern Recognition & Machine Learning
Why use the EM Algorithm for PCA?

- For large datasets, can be more computationally efficient than an eigendecomposition or SVD
- Regularization: can put priors on model parameters, do Bayesian model order selection, etc.
- Cleanly handles cases where some entries of the data matrix are unobserved or missing (e.g., movie ratings)
- Generalizes to other models where there is no closed form for the maximum likelihood estimates (e.g., factor analysis)

Probabilistic PCA or Factor Analysis

- Probabilistic PCA models all rotations of the input data equally well (are basis vectors meaningful?)
- Factor analysis models all element-wise rescalings of the input data equally well (better when varying units)
Factor Analysis Example

Features of Cars in 2004

- Suggested retail price in USD
- Price to dealer in USD
- Engine size in liters
- Num. cylinders
- Horsepower
- City MPG
- Highway MPG
- Weight in pounds
- Wheelbase in inches
- Length in inches
- Width in inches