Machine learning II

Bayesian data modeling

Model based machine learning

General setup of model based ML:

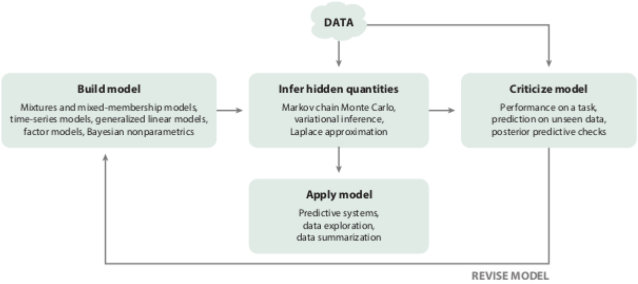


Fig. from: David M. Blei, *Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models*, Annu. Rev. Stat. Appl. 2014. 1:20332

Statistical modeling

Have already seen different classical models/algorithms:

- K-means clustering
- Linear regression

Statistical modeling

Have already seen different classical models/algorithms:

- K-means clustering
- Linear regression

In both cases, we could find a statistical interpretation:

- Data generated from latent classes, i.e. mixture model

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mu_k)\mathcal{N}(\mathbf{x}|\mu_k, \sigma_\epsilon^2)$$

- Noisy observation of function, i.e.

$$p(t|\mathbf{x}) = \mathcal{N}(t|y(\mathbf{x}), \sigma_\epsilon^2)$$

where $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$.

Statistical modeling

Two different types of models

- **Discriminative**:
  - Model relation between input $\mathbf{x}$ and target output t
  - Requires labelled training data, i.e. *supervised learning*
- **Generative**:
  - Model dependencies within observed data $\mathbf{x}$

$$
\begin{aligned}
p(\mathbf{x}) &= \int p(\mathbf{x}, \mathbf{z})\, d\mathbf{z} \\
&= \int p(\mathbf{z}) p(\mathbf{x}|\mathbf{z})\, d\mathbf{z}
\end{aligned}
$$

    by explaining them via *latent variables* $\mathbf{z}$
  - Unlabelled data suffice, i.e. *unsupervised learning*

Bayesian modeling

- Model completed by prior $p(\theta)$ on parameters, i.e.

  $$p(\mathbf{x}, \theta) = p(\mathbf{x}|\theta)p(\theta)$$

  Note: **No** difference between parameters and latent variables!
- Can be used to generate artificial data, either conditionally (discriminative) or unconditionally (generative), i.e.

  $$p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$$

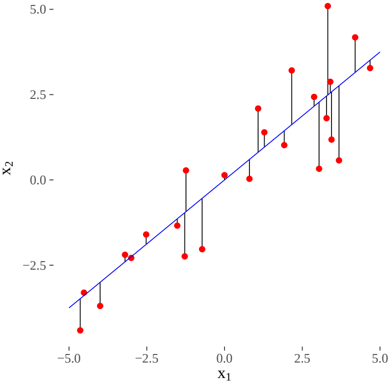
- Inference based on Bayes rule, i.e. posterior distribution

  $$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)p(\theta)$$

  Principled, mechanical, intractable ...

From linear regression ...

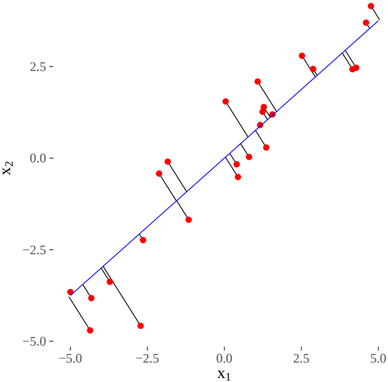
Linear regression (in 2D):



- Linear relation between x_1 and x_2
- x_2 is observed with noise, x_1 is noiseless
- Conditional model $p(x_2|x_1)$

Principle component analysis

Principal component analysis (in 2D):



- Linear relation between x_1 and x_2
- Both x_1 and x_2 are observed with noise
- Generative model $p(x_1, x_2)$

Probabilistic principal component analysis

- Data $\mathbf{x} \in \mathbb{R}^D$
- Continuous latent variable $\mathbf{z} \in \mathbb{R}^Q$
- Generative model for data:

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathbf{x} &\sim \mathcal{N}(\mathbf{W}\mathbf{z}, \sigma^2 \mathbf{I}) \end{aligned}$$

  with parameters $\mathbf{W}$ and σ^2

- Data distribution is Gaussian with low-rank covariance matrix

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\, d\mathbf{z} = \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I})$$

...back to K-means

- Data $\mathbf{x} \in \mathbb{R}^D$
- Discrete latent variable $c \in \{1, \ldots, K\}$
- Generative model for data:

$$\begin{aligned} c &\sim \mathcal{C}\text{ategorical}(\theta) \\ \mathbf{x} &\sim \mathcal{N}(\mu_c, \sigma^2 \mathbf{I}) \end{aligned}$$

  with parameters $\theta, \{\mu_c\}_{c=1}^K$ and σ^2

- Data distribution is mixture of Gaussians

$$p(\mathbf{x}) = \sum_{c=1}^{K} \theta_c \mathcal{N}(\mathbf{x} | \mu_c, \sigma^2 \mathbf{I})$$

…back to K-means

- Recode latent class c as

$$\mathbf{z}^c \in \mathbb{R}^K \text{ with } z_i^c = \begin{cases} 1 & \text{if } c = i \\ 0 & \text{otherwise} \end{cases}$$

- Collect means $\{\mu_c\}_{c=1}^K$ in weight matrix

$$\mathbf{W} = (\mu_1, \ldots, \mu_K)$$

- Then, we note that

$$\begin{array}{rrcl} & \mathbf{x} & \sim & \mathcal{N}(\mu_c, \sigma^2 \mathbf{I}) \\ \text{is the same as} & \mathbf{x} & \sim & \mathcal{N}(\mathbf{W}\mathbf{z}^c, \sigma^2 \mathbf{I}) \end{array}$$

Bayesian modeling

Are PPCA and K-means the same model?

- Same sampling distribution $p(\mathbf{x}|\mathbf{z})$ ✓
- But very different prior $p(\mathbf{z})$ ✗

PPCA	$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
K-means	$\mathbf{z}^c$ with $c \sim \mathcal{C}\text{ategorical}(\theta)$

What is a Bayesian model?

- Model consists of **both** prior (for latent variables and parameters) **and** sampling distribution/likelihood
- Model defines generative story for data

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}, \theta) d\mathbf{z} d\theta$$

Good model should be able to generate plausible data!

Other models

- Sparse linear regression
  - Sparsity prior to select relevant inputs
  - Examples: LASSO, Horseshoe prior
- Factor analysis
  - Generalization of PPCA

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \mu, \mathbf{\Psi})$$

  where $\mathbf{\Psi}$ is a diagonal matrix
- Independent component analysis
  - Developed for blind source separation
  - Generalization of PCA

$$p(\mathbf{z}) = \prod_m p(z_m)$$

  with non-Gaussian marginal distributions
- Gaussian mixture models
  - Generalization of K-means clustering
  - Flexible models for multi-modal data distributions

Summary

- Discrimative/generative model vs supervised/unsupervised learning
- Probabilistic models for data distribution:
  - Sparse linear regression
  - Latent variable models
    - Discrete: Mixture models
    - Continuous: PCA, ICA, manifold models, ...
- Bayesian approach:
  - Uncertainty estimates
  - Missing data and prediction of new data
  - Selection of model complexity
- Current research:
  - Probabilistic interpretation of deep learning