Machine learning II
Bayesian data modeling
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

\[
p(x) = \sum_{k=1}^{K} p(\mu_k)\mathcal{N}(x|\mu_k, \sigma^2_{\epsilon})
\]

▶ Noisy observation of function, i.e.

\[
p(t|x) = \mathcal{N}(t|y(x), \sigma^2_{\epsilon})
\]

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

Two different types of models

- **Discriminative:**
  - Model relation between input $x$ and target output $t$
  - Requires labelled training data, i.e. *supervised learning*

- **Generative:**
  - Model dependencies within observed data $x$
    
    
    $$
    p(x) = \int p(x, z) \, dz
    $$
    
    $$
    = \int p(z)p(x|z) \, dz
    $$

    by explaining them via *latent variables* $z$
  - Unlabelled data suffice, i.e. *unsupervised learning*
Bayesian modeling

- Model completed by prior $p(\theta)$ on parameters, i.e.
  \[ p(x, \theta) = p(x|\theta)p(\theta) \]

  Note: **No** difference between parameters and latent variables!

- Can be used to generate artificial data, either conditionally (discriminative) or unconditionally (generative), i.e.
  \[ p(x) = \int p(x|\theta)p(\theta)d\theta \]

- Inference based on Bayes rule, i.e. posterior distribution
  \[ p(\theta|x) \propto p(x|\theta)p(\theta) \]

  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:

  $$\mathbf{z} \sim \mathcal{N}(0, I)$$
  $$\mathbf{x} \sim \mathcal{N}(\mathbf{Wz}, \sigma^2 I)$$

  with parameters $\mathbf{W}$ and $\sigma^2$
- Data distribution is Gaussian with low-rank covariance matrix

  $$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,d\mathbf{z} = \mathcal{N}(0, \mathbf{WW}^T + \sigma^2 I)$$
... back to K-means

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

  $$c \sim \text{Categorical}(\theta)$$

  $$\mathbf{x} \sim \mathcal{N}(\mu_c, \sigma^2 I)$$

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

- Data distribution is mixture of Gaussians

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

- Recode latent class $c$ as

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

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

$$W = (\mu_1, \ldots, \mu_K)$$

- Then, we note that

$$x \sim \mathcal{N}(\mu_c, \sigma^2 I)$$

is the same as

$$x \sim \mathcal{N}(Wz^c, \sigma^2 I)$$
Are PPCA and K-means the same model?

- Same sampling distribution $p(x|z)$ ✓
- But very different prior $p(z)$ ×
  - **PPCA** $z \sim \mathcal{N}(0, I)$
  - **K-means** $z^c$ with $c \sim \text{Categorical}(\theta)$

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

$$p(x) = \int p(x, z, \theta) dz d\theta$$

Good model should be able to generate plausible data!
Case study

Investigate data about putting success of professional golfers

Will use *Stan* to model these data!
Probabilistic programming language **Stan:**
- Specify statistical model as computer program
- Inference is done automatically:
  - Model compiled to C++
  - Automatic differentiation for efficient sampling
- Posterior represented via samples
Linear regression

Stan code for linear regression:

data {
    int<lower=0> N; // Number of training data
    int<lower=0> D; // Number of regressors
    vector[D] X[N]; // Regressor inputs
    real t[N]; // Target outputs
    real<lower=0> alpha; // Hyperparameters
    real<lower=0> beta;
    real<lower=0> tau0;
}
}

parameters {
    vector[D] w; // Regression weights
    real<lower=0> tau_eps; // Noise precision
}
}

model {
    // Priors
    tau_eps ~ gamma(alpha, beta);
    w ~ normal(0, sqrt(1/tau0));
    // Likelihood
    for (n in 1:N) {
        real y;
        y = w’ * X[n]; // dot product w^T x_n
        t[n] ~ normal(y, sqrt(1/tau_eps));
    }
}

- Program specifies \( \ln p(\text{data, parameters}) \)
- Samples/optimizes/approximates \( p(\text{parameters} \mid \text{data}) \)
Golf putting

How would you model Golf putting data?

- What about a linear model?
- What about a quadratic model?
How would you model Golf putting data?

- What about a linear model?
- What about a quadratic model?
  - Note: Probabilities are bounded in $[0, 1]$
  - Think about predictions at extreme distances

Is your model reasonable?
Golf putting

Logistic regression

- Linear model for classification/probabilistic prediction

\[ P(Y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}) \]

\[ y \sim \text{Bernoulli}(\sigma(\mathbf{w}^T \mathbf{x})) \]

- Generalized linear model with logit *link function*

\[ \text{logit}(\theta) = \log \left( \frac{\theta}{1 - \theta} \right) \]

\[ \sigma(\alpha) = \frac{1}{1 + e^{-\alpha}} = \text{logit}^{-1}(\alpha) \]
Golf putting

Logistic regression

pred_prob

prob

0.95
Assume that angular precision \( \theta \) is random, e.g. \( \theta \sim \mathcal{N}(0, \sigma) \)

Translates into statistical model for observations

\[
P(\text{Success prob. from distance } x) = P(|\theta| < \theta_0)
= 2\Phi\left(\frac{1}{\sigma} \arcsin\left(\frac{R-r}{x}\right)\right) - 1
\]
Summary

- Discriminative/generative model vs supervised/unsupervised learning
- Probabilistic models for data distribution:
  - Generalized 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