STATS 235: Modern Data Analysis
Introduction

Babak Shahbaba

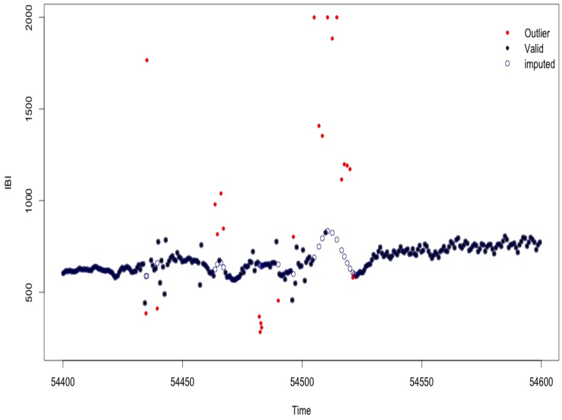
Department of Statistics, UCI

Statistical methods in machine learning

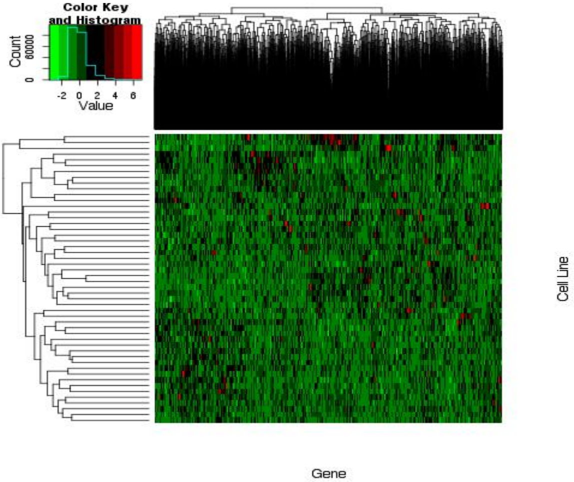
- In this course, we focus on complex and/or high dimensional problems
- We discuss statistical methods designed for such problems
- Our overall objective is to use these statistical methods to make decisions under uncertainty
- Typically, our decisions are in the form of predictions
- To achieve this objective, we need to learn from the data: detect pattern, discover relationships, and possibly reduce dimensionality and complexity along the way
- In this lecture, we discuss some motivating examples and review some introductory concepts

Motivating Examples

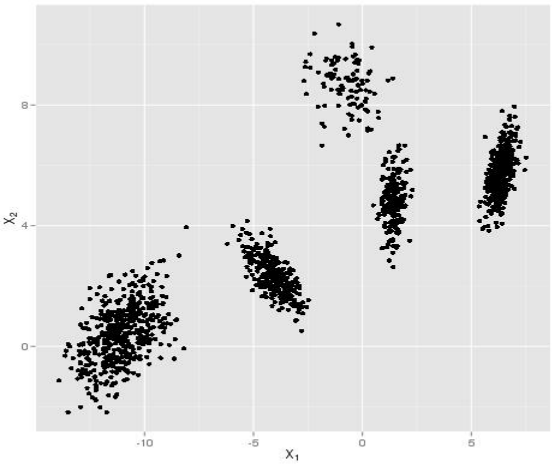
Automatic data cleaning and processing



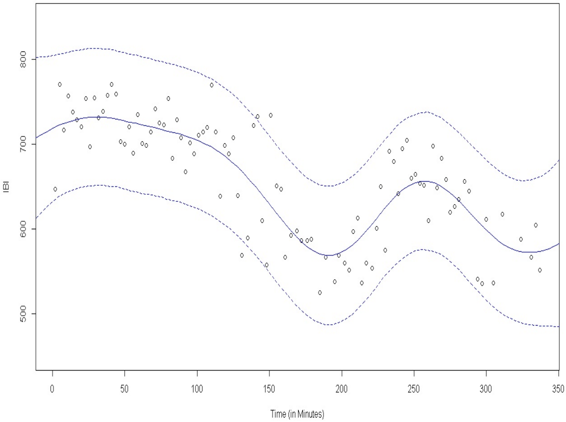
High throughput biological studies



Clustering objects



Regression



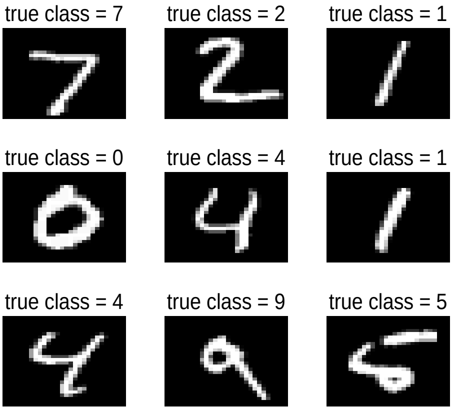


Fig1.5a in Murphy (2012)

Document classification

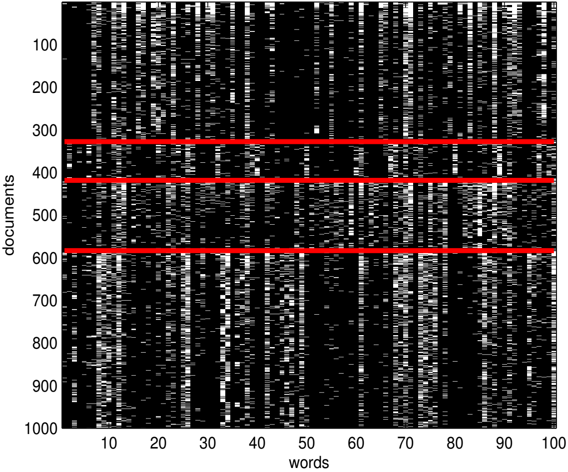
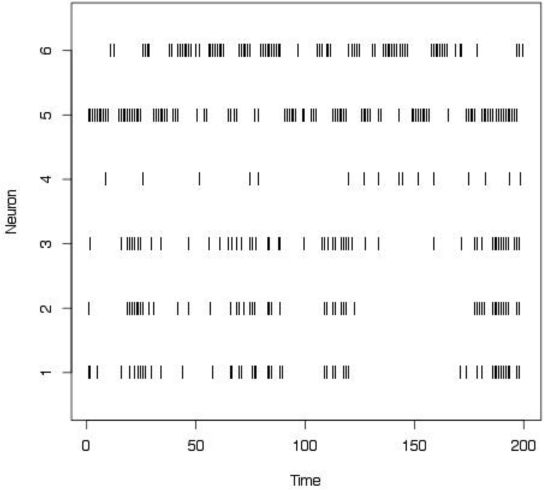


Fig1.2 in Murphy (2012)

Neural coding



Discovering biological networks

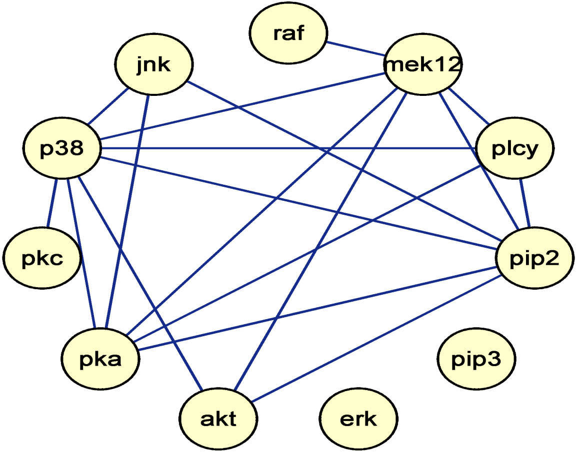


Fig1.11 in Murphy (2012)

Some Preliminary Concepts

Supervised vs. unsupervised learning

- Learning problems discussed in machine learning are divided into two main categories
  - Supervised learning: our objective is to find (learn) a mapping from a set of inputs (predictors, attributes, covariates, features), x, to outputs (response, target, outcome), y.
  - Unsupervised learning: there are no clear, well-defined outputs; our objective is to discover interesting patterns, relationships, and structures in a set of inputs, x.
- Semi-supervised and reinforcement learning are two of less commonly used learning methods.

Supervised learning

- Regression
  - Continuous response variables: $y \in \mathbb{R}$
  - Forecasting
  - Longitudinal analysis
  - Time series analysis
  - Spatio-temporal analysis
- Classification
  - Categorical response variable: $y \in \{1, \ldots, C\}$
  - Document classification
  - Image classification
  - Face detection and recognition

Discriminative vs. generative models

- Discriminative classification models
  - We model $P(y|x)$, $y \in \{1, \ldots, C\}$, and use it to predict the class given inputs.
  - Tends to be less sensitive to outliers
  - Possible to use arbitrary preprocessing of inputs
- Generative classification models
  - We model $P(x|y)$, $y \in \{1, \ldots, C\}$, and use Bayes' rule to find $P(y|x)$ in order to make predictions given inputs.
  - Easy to fit
  - Can handle missing features and unlabeled data
  - If the assumed distribution of inputs is correct, they tend to perform better since they use more information to estimate parameters.

Unsupervised learning

- Density estimation
  - Finite mixture models
  - Infinite mixture models
- Clustering
  - K-means clustering
  - Hierarchical clustering
  - Finite mixture models
- Dimensionality reduction
  - Principal component analysis (PCA)
  - Factor analysis (FA)
  - Independent component analysis (ICA)

Parametric vs. nonparametric

- In general, supervised learning involves finding $P(y|x)$, whereas, unsupervised learning involves finding $P(x)$.
- More specifically, we use $P(y|x, \theta)$ and $P(x|\theta)$, where θ represent all the model parameters.
- Parametric models: Number of parameters is fixed (finite).
- Nonparametric models: Number of parameters grows (possibly to infinity) with the amount of data.

Some Common Parametric Models

Exponential family of distributions

- Very often, we assume simple distributional forms (e.g., normal, binomial, Poisson) that are members of the exponential family.

- A single parameter distributional form belongs to the exponential family if the distribution has the following form

$$P(y|\theta) = \frac{1}{z(\theta)} h(y) \exp[g(\theta) s(y)]$$

or

$$P(y|\theta) = h(y) \exp[g(\theta) s(y) - c(\theta)]$$

- P is the density function for continuous random variables and probability mass function for discrete variables.

- $z(\theta)$ is called the "partition function."

Poisson distribution

- For example, for Poisson distributions, we have

$$\begin{aligned} P(y|\theta) &= e^{-\theta}\theta^{y}/y! \\ &= \frac{1}{y!}\exp\{\log(\theta)y - \theta\} \end{aligned}$$

- Here,

$$\begin{aligned} g(\theta) &= \log(\theta) \\ s(y) &= y \\ c(\theta) &= \theta \\ h(y) &= \frac{1}{y!} \end{aligned}$$

Sufficiency in exponential family

- For a sample of n independent observations, $y = (y_1, y_2, ..., y_n)$, we have

$$P(y|\theta) = \prod h_i(y_i) \times \exp\{\sum g_i(\theta)s_i(y_i) - \sum c_i(\theta)\}$$

- If the observations are identically distributed, we can drop the index i, and present the exponential family as

$$P(y|\theta) = h(y)\exp\{g(\theta)s(y) - c(\theta)\}$$

- $s(y)$ is the "sufficient statistic" for θ.

- $\phi = g(\theta)$ is called the "natural parameter."

Multiparameter exponential family

- In general, a distribution in exponential family can have multiple parameters.

- In this case, we define the distribution in term of $g^T(\theta)s(y)$,

$$P(y|\theta) = h(y)\exp\{\sum_{k=1}^{K} g_k(\theta)s_k(y) - c(\theta)\}$$

where $s = (s_1, s_2, \ldots, s_k)$ is sufficient for θ.

- Note that while the dimension of s, which is K, is usually the same as the dimension of θ, this does not have to be the case in general.

- Also note that g and s are not unique.

Normal distribution

- Let's consider a normal distribution with unknown mean and unknown variance.

$$
\begin{aligned}
P(y|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}} \exp\{\frac{-(y-\mu)^2}{2\sigma^2} - \frac{\log(\sigma^2)}{2}\} \\
&= \frac{1}{\sqrt{2\pi}} \exp\{-\frac{y^2}{2\sigma^2} + \frac{\mu y}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{\log(\sigma^2)}{2}\}
\end{aligned}
$$

- Here,

$$
\begin{aligned}
g(\mu, \sigma^2) &= (\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}) \\
s(y) &= (-\frac{y^2}{2}, y) \\
c(\mu, \sigma^2) &= \frac{\mu^2}{2\sigma^2} + \frac{\log(\sigma^2)}{2} \\
h(y) &= 1/\sqrt{2\pi}
\end{aligned}
$$

Generalized linear models

- In generalized linear models, where linear regression is a special case, we assume that the response variable, y, has a distribution in the exponential family.

- In these models an element of the natural parameter, $g(\theta)$, is a function of $E(y|\theta)$.

- We usually use this function as a link between μ and inputs,

$$g(\mu_i) = x_i^\top \beta$$

- We refer to $g(\mu)$ as the canonical link and use it to rewrite the model in terms of regression parameters β.

Poisson regression model

- In the Poisson model discussed above, $\theta = \mu$. Therefore,

$$
\begin{aligned}
P(y|\mu) &= e^{-\mu}\mu^{y}/y! \\
&= \frac{1}{y!}\exp\{\log(\mu)y - \mu\}
\end{aligned}
$$

- In this case, the canonical link function is

$$
g(\mu_i) = \log(\mu_i) = x_i^\top \beta
$$

- Therefore,

$$
P(y_i|x_i, \beta) = \frac{1}{y_i!}\exp\{(x_i^\top \beta)y_i - \exp(x_i^\top \beta)\}
$$

- For n iid observations,

$$
P(y_1, \ldots, y_n|x_i, \ldots, x_n, \beta) = \frac{1}{\prod y_i!}\exp\left\{\sum[(x_i^\top \beta)y_i - \exp(x_i^\top \beta)]\right\}
$$

Ising model

- Another model with an exponential family distribution is the Ising model commonly used in statistical physics and graphical models (see MacKay).

- An Ising model is an array of spins (denoted as ± 1) that are magnetically coupled to each other.
  - Ferromagnetic model: if one spin is $+1$, it is energetically favorable for its immediate neighbors to be $+1$.
  - Antiferromagnetic model: if one spin is $+1$, it is energetically favorable for its immediate neighbors to be -1.

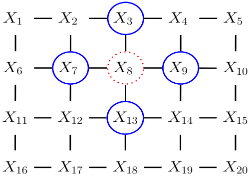


Fig19.1b in Murphy (2012)

Ising model

- Two spins i and j are neighbors: $(i,j) \in \mathcal{N}$.

- The energy of the a specific configuration, X, is given by Hamiltonian function,

$$E(X, \alpha, \beta) = -\frac{1}{2}\sum_{i,j} \beta_{ij} x_i x_j - \sum_i \alpha_i x_i$$

where

  - β_{ij} represents the coupling (interaction) between two neighboring spins such that $\beta_{ij} = 0$ if $(i,j) \notin \mathcal{N}$.

  - α_i represents the external magnetic field on spin i.

- The probability of any specific configuration, X, is given by the Boltzmann distribution,

$$
\begin{aligned}
P(X|\alpha, \beta, T) &= \frac{1}{z(\alpha, \beta, T)} \exp\{-\frac{1}{K_B T} E(X, \alpha, \beta)\} \\
&= \frac{1}{z(\alpha, \beta, T)} \exp\{\frac{1}{K_B T}[\frac{1}{2} \sum_{i,j} \beta_{ij} x_i x_j + \sum_i \alpha_i x_i]\}
\end{aligned}
$$

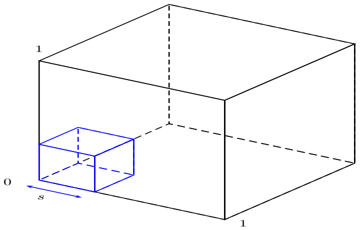
where,

$$
\begin{aligned}
z(\alpha, \beta, T) &= \sum_x \exp\{-\frac{1}{K_B T} E(X, \alpha, \beta)\} \\
P(X_i = +1|.) &= \frac{1}{1 + \exp(-\frac{2}{K_B T} \sum_j \beta_{ij} x_j + \alpha_i)}
\end{aligned}
$$

- Here, T is the temperature and K_B is Boltzmann's constant.

Some Modeling Challenges

- Curse of dimensionality (Bellman, 1961) refers to challenges imposed by high dimensional data.
- These challenges are mainly due to sparsity.
- Hastie et. al. (2009) have demonstrated this using a simple example.



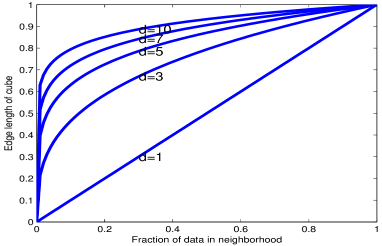


Fig1.6 in Murphy (2012)

Curse of dimensionality

- Suppose inputs are uniformly distributed in d-dimensional unit hypercube.
- We want to estimate the output (e.g., classification) at each point, x_0, using a fraction r (e.g., 0.01) of inputs in a hypercubical neighborhood of x_0.
- The expected edge length is $e_p(r) = r^{1/d}$.
- When $d = 2$, $e_2(0.01) = 0.1$.
- When $d = 10$, $e_{10}(0.01) = 0.63$.

Overfitting

- Overfitting is a common challenge in applying machine learning methods.
- It refers to situations when a model performance well on the observed data, but performs poorly on future observations.
- This is mainly due to the fact that many machine learning methods can lead to arbitrarily complex models.
- As a result, these models identify patterns peculiar to the observed data but not generalizable to the whole population.

Occam's razor

- We will talk about techniques for controlling complexity throughout this course.
- In general, it is recommended to use more complex models only when they result in substantial (i.e., statistically significant) improvement in performance (i.e, substantial decrease in deviance).
- The above principle is widely known as Occam's razor stating that "entities should not be multiplied beyond necessity", or in simple words: "everything equal, we should use the simplest solution".
- Ideally, we prefer to use complex models that include simpler models as special cases and have an intrinsic complexity-controlling mechanism.

Model comparison and model selection

- As we will see later, model selection is properly defined within the decision theory framework.
- Decision theory is however an easy concept that is hard to implement.
- An essential element of a decision making problem is the specification of a proper loss function.
- For regression models, the squared error loss function is commonly used,

$$L(y, \hat{y}) = (y - \hat{y})^2$$

where $\hat{y}$ is the estimated value of unknown (e.g., future) y based on our model.

Model comparison and model selection

- For predictive models, we can define our goal as finding the model with the lowest expected loss, in this case the lowest expected prediction error $EPE = E[L(y, \hat{y})]$.

$$
\begin{aligned}
EPE &= E(y^2) + E(\hat{y}^2) - 2E(y\hat{y}) \\
&= \mathrm{Var}(y) + E(y)^2 + \mathrm{Var}(\hat{y}) + E(\hat{y})^2 - 2E(y)E(\hat{y}) \\
&= \mathrm{Var}(y) + (E(y) - E(\hat{y}))^2 + \mathrm{Var}(\hat{y}) \\
&= \mathrm{Var}(y) + \mathrm{Bias}^2(\hat{y}) + \mathrm{Var}(\hat{y})
\end{aligned}
$$

- Note that future observations y are independent of $\hat{y}$.

Bias-variance tradeoff

- In the above derivation of EPE, the first term, $\mathrm{Var}(y)$, reflects the random variation of the response variable regardless of what model we use.
- Therefore, only the last two terms depend on our model for $\hat{y}$; so we should try to minimize these two terms.
- In general, there is a tradeoff between bias and variance: complex models tend to have lower bias and higher variance, whereas simple models tend to have higher bias and lower variance.

Course outline

The remaining parts of this course are organized as follows:

- An overview of decision theory and learning frameworks
  - Frequentist approach
  - Bayesian statistics
  - Information theory
- Unsupervised learning
  - Dimensionality reduction
  - Clustering
  - Mixture models for density estimation
- Supervise learning
  - Regression
  - Classification
- Graphical models