COMS 4721: Machine Learning for Data Science
Lecture 1, 1/17/2017

Prof. John Paisley

Department of Electrical Engineering
& Data Science Institute

Columbia University

Overview

This class will cover model-based techniques for extracting information from data with an end-task in mind. Such tasks include:

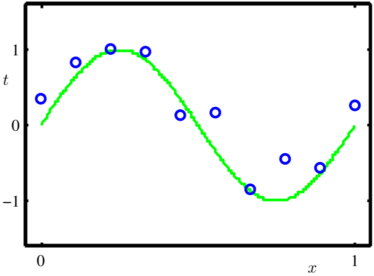
1 - predicting an unknown "output" given its corresponding "input"
2 - uncovering information within the data to better understand it
3 - data-driven recommendation, grouping, classification, ranking, etc.

There are a few ways we can divide up the material as we go along, e.g.,

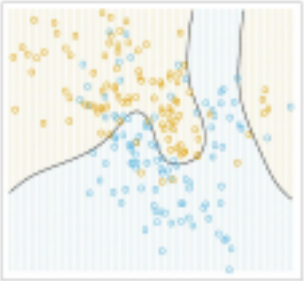
supervised learning	|	unsupervised learning
probabilistic models	|	non-probabilistic models
modeling approach	|	optimization techniques

We'll adopt the first method and work in the second two along the way.

OVERVIEW: SUPERVISED LEARNING



(a) Regression



(b) Classification

Regression: Using set of inputs, predict real-valued output.

Classification: Using set of inputs, predict a discrete label (aka class).

Example Classification Problem

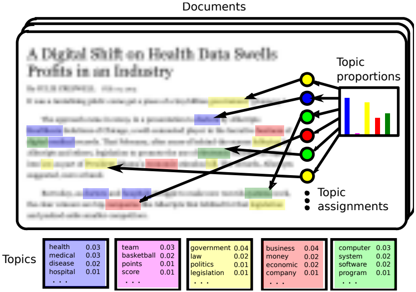
Given a set of inputs characterizing an item, assign it a label.

Is this spam?

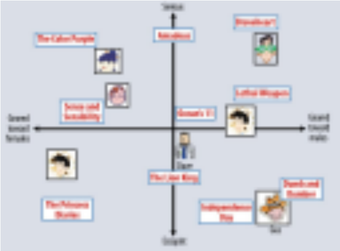
```
hi everyone,

i saw that close to my hotel there is a pub with bowling
(it’s on market between 9th and 10th avenue).  meet
there at 8:30?
```

What about this?

```
Enter for a chance to win a trip to Universal Orlando to
celebrate the arrival of Dr. Seuss’s The Lorax on Movies
On Demand on August 21st!  Click here now!
```



(c) topic modeling



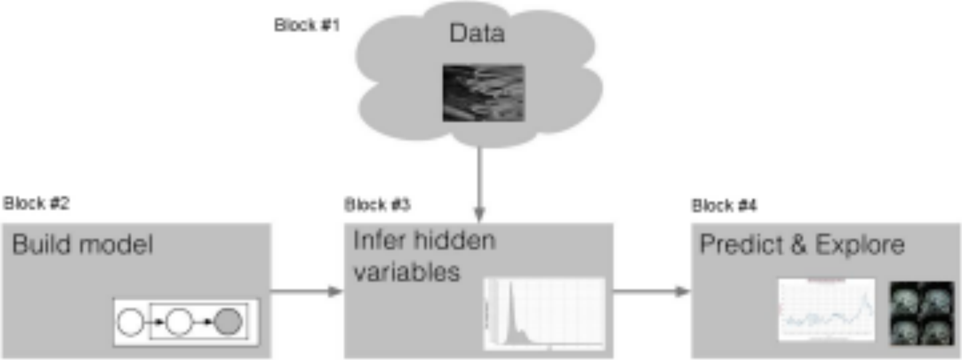
(d) recommendations[1]

With unsupervised learning our goal is often to uncover structure in the data. This helps with predictions, recommendations, efficient data exploration.

[1] Figure from Koren, Y., Robert B., and Volinsky, C.. "Matrix factorization techniques for recommender systems." Computer 42.8 (2009): 30-37.

Goal: Learn the dominant topics from a set of news articles.

Data Modeling



- Supervised vs. unsupervised: Blocks #1 and #4
- Probabilistic vs. non-probabilistic: Primarily Block #2 (Some Block #3)
- Model development (Block #2) vs. Optimization techniques (Block #3)

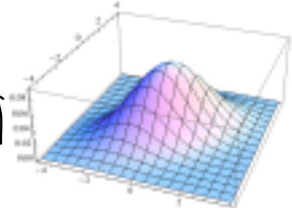
GAUSSIAN DISTRIBUTION (MULTIVARIATE)

Gaussian density in d dimensions

d=2

- Block #1: Data $x_1, \ldots, x_n$. Each $x_i \in \mathbb{R}^d$
- Block #2: An i.i.d. Gaussian model
- Block #3: Maximum likelihood
- Block #4: Leave undefined

value of f



The density function is

Probability of x given parameters μ & Σ

i/p vector

non-negative, integrate to 1.

$$p(x|\mu,\Sigma) := \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\underbrace{\Sigma}_{\text{positive definite}})}} \exp\left(-\frac{1}{2}\underbrace{(x-\mu)^T\Sigma^{-1}(x-\mu)}_{\text{quadratic function}}\right)$$

[We evaluate this function and that gives a density at x using a gaussian with mean μ & covariance Σ]

The central moments are:

$$\mathbb{E}[x] = \int_{\mathbb{R}^d} x\, p(x|\mu,\Sigma)dx = \mu,$$

subtract mean of x from x pt.

and then evaluate expectation of this outer product

equal to sigma

$$\text{Cov}(x) = \mathbb{E}[(\overbrace{x-\mathbb{E}[x]}^{})(x-\mathbb{E}[x])^T] = \mathbb{E}[xx^T] - \mathbb{E}[x]\mathbb{E}[x]^T = \Sigma.$$

↳ spread of distribution in various dimensions, as well as the correlation between those dimensions

Block #2: A Probabilistic Model

simply a set of probability distributions on our data.

Probabilistic Models

Where data X, we have to define some probability distribution on X. That distribution is going to take parameters θ

- A *probabilistic model* is a set of probability distributions, $p(x|\theta)$.
- We pick the *distribution family* $p(\cdot)$, but don't know the parameter θ.

No matter what value we set θ to, we are always working within the same distribution family.

Example: Model data with a Gaussian distribution $p(x|\theta)$, $\theta = \{\mu, \Sigma\}$.

this density is a multi-variate Gaussian & takes parameters μ & Σ.

For all parameters μ & Σ, we are working with the same family.

(often made)

The i.i.d. assumption

Assume data is *independent and identically distributed (iid)*. This is written

$$x_i \stackrel{iid}{\sim} p(x|\theta), \quad i = 1, \ldots, n.$$

every single data point is simply independent of every other and it has the same distribution family.

Writing the density as $p(x|\theta)$, then the *joint* density decomposes as

This assumption allows us to write our joint density this way

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

Joint density (probability) of all n observations is simply the product of the prob. of each of them individually.

Probability of n d-dimensional vectors coming from a distribution with parameter θ

BLOCK #3: MAXIMUM LIKELIHOOD ESTIMATION

(define an objective function).

of data & unknown model parameters

Maximum Likelihood approach

(common approach with probabilistic models.)

We now need to find θ. *Maximum likelihood* seeks the value of θ that maximizes the likelihood function:

For a particular value of Θ, it tells us what is the likelihood of data coming from that model.
What is the prob. of observations that we have given that particular value of Θ.

[Find Θ that maximises joint likelihood function.]

$$\hat{\theta}_{\mathrm{ML}} := \arg\max_{\theta}\; p(x_1, \ldots, x_n|\theta),$$

prob. distribution on our data [joint likelihood]

This value best explains the data according to the chosen distribution family.

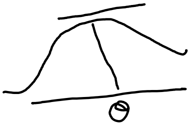
we want to find the value of Θ that says this dataset is most probable from a m.v. Gaussian having this parameterization.

[In a sense we are fitting the data.

$\hat{\Theta}_{ML} \rightarrow$ maximum likelihood estimate

Maximum Likelihood equation

The analytic criterion for this maximum likelihood estimator is:



$$\nabla_{\theta} \prod_{i=1}^{n} p(x_i|\theta) = 0.$$

Maximum of a function is the pt at which gradient $= 0$.

Simply put, the maximum is at a peak. There is no “upward” direction.

BLOCK #3: LOGARITHM TRICK (This technique is problem specific)

(When using prob. distributions in an iid setting.)

(log doesn't change locations of maximum & minimum)

Logarithm trick

Calculating $\nabla_\theta \prod_{i=1}^{n} p(x_i|\theta)$ can be complicated. We use the fact that the logarithm is monotonically increasing on $\mathbb{R}_+$, and the equality

$$\ln\Big(\underbrace{\prod_i f_i}_{\text{product}}\Big) = \underbrace{\sum_i \ln(f_i)}_{\text{sum}}.$$

Consequence: Taking the logarithm does not change the *location* of a maximum or minimum:

$$\max_y \ \ln g(y) \neq \max_y \ g(y) \qquad \text{The } \textit{value} \text{ changes.}$$

($g \neq \ln g$)

$$\arg\max_y \ \ln g(y) = \arg\max_y \ g(y) \qquad \text{The } \textit{location} \text{ does not change.}$$

which we care about

Maximum likelihood and the logarithm trick

(iid assumption)   product   sum

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta) = \arg\max_{\theta} \ln\left(\prod_{i=1}^{n} p(x_i|\theta)\right) = \arg\max_{\theta} \sum_{i=1}^{n} \ln p(x_i|\theta)$$

[we work with this function.]

To then solve for $\hat{\theta}_{\text{ML}}$, find

$$\nabla_{\theta} \sum_{i=1}^{n} \ln p(x_i|\theta) = \sum_{i=1}^{n} \nabla_{\theta} \ln p(x_i|\theta) = 0.$$

[gradient at $\theta = 0$]

Gradient of each individual likelihood over each data point

Depending on the choice of the model, we will be able to solve this

1. analytically (via a simple set of equations) → in this case
2. numerically (via an iterative algorithm using different equations)
3. approximately (typically when #2 converges to a local optimal solution)

in complicated models we can't take the derivative & set it = 0.

Block #2: Multivariate Gaussian data model

Model: Set of all Gaussians on $\mathbb{R}^d$ with unknown mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{S}^d_{++}$ (positive definite $d \times d$ matrix).

We assume that $x_1, \ldots, x_n$ are i.i.d. $p(x|\mu, \Sigma)$, written $x_i \stackrel{iid}{\sim} p(x|\mu, \Sigma)$.

Block #3: Maximum likelihood solution

in case it is maximizing the f^n because we want it to be more probable.

We have to solve the equation

$$\sum_{i=1}^{n} \nabla_{(\mu,\Sigma)} \ln p(x_i|\mu, \Sigma) = 0$$

for μ and Σ. (Try doing this without the log to appreciate it's usefulness.)

EXAMPLE: GAUSSIAN MEAN MLE

First take the gradient with respect to μ.

$$0 = \nabla_\mu \sum_{i=1}^{n} \ln \overbrace{\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)}^{\text{multi-variate Gaussian evaluated at each observation.}}$$

$$= \nabla_\mu \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi)^d|\Sigma| - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

[Sum of log of each of these 2 terms]

?

$$= -\frac{1}{2}\sum_{i=1}^{n} \nabla_\mu \left(x_i^T\Sigma^{-1}x_i - 2\mu^T\Sigma^{-1}x_i + \mu^T\Sigma^{-1}\mu\right) = -\Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu)$$

Set = $\begin{bmatrix} 0 \\ \vdots \\ \vdots \end{bmatrix}$ $\vec{0}$ vector (since it is not a scalar)

Gradient of the log of joint likelihood with respect to mean vector μ.

Since Σ is positive definite, the only solution is

$$\sum_{i=1}^{n}(x_i-\mu) = 0 \qquad \Rightarrow \qquad \hat{\mu}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Since this solution is independent of Σ, it doesn't depend on $\hat{\Sigma}_{\text{ML}}$.

EXAMPLE: GAUSSIAN COVARIANCE MLE

Now take the gradient with respect to Σ. (dxd matrix)

matrix of 0s
dxd

$$0 = \nabla_\Sigma \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi)^d|\Sigma| - \underline{\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}$$

$$= -\frac{n}{2}\nabla_\Sigma \ln|\Sigma| - \frac{1}{2}\nabla_\Sigma \text{trace}\Big(\Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\Big)$$

?

?

$$= -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-2}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$$

(Function of μ & Σ)

= equating to matrix of 0s

Solving for Σ and plugging in $\mu = \hat{\mu}_{\text{ML}}$,

$$\hat{\Sigma}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{\text{ML}})(x_i-\hat{\mu}_{\text{ML}})^T.$$

} empirical avg. of our outer products of our data where we have substracted of max. likelihood mean.

EXAMPLE: GAUSSIAN MLE (SUMMARY)

So if we have data $x_1, \ldots, x_n$ in $\mathbb{R}^d$ that we hypothesize is i.i.d. Gaussian, the maximum likelihood values of the mean and covariance matrix are

(empirical covariance of data where we subtract off max.-likelihood mean.)

$$\hat{\mu}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\Sigma}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\text{ML}})(x_i - \hat{\mu}_{\text{ML}})^T.$$

Are we done? There are many assumptions/issues with this approach that makes finding the "best" parameter values not a complete victory.

- We made a model assumption (multivariate Gaussian).
- We made an i.i.d. assumption. (sequential information may be present) → why would it not be?
- We assumed that maximizing the likelihood is the best thing to do.

Comment: We often use θ_{ML} to make predictions about x_{new} (Block #4).
How does θ_{ML} generalize to x_{new}?
If $x_{1:n}$ don't "capture the space" well, θ_{ML} can *overfit* the data.

The problem was solved for a function that we chose.