ISTA 421 + INFO 521
Introduction to Machine Learning

Lecture 19: Bayesian Classification

Clay Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

1 November 2017

1

Spring Courses Advertisement

- ISTA 450 / INFO 550 – Introduction to AI
  - Discrete problem solving, search, adversarial search (game search), constraint satisfaction, logical inference, decision-making under uncertainty, reinforcement learning
- ISTA 457 / INFO 557 – Neural Networks
  - Steven Bethard – not yet in RCS, but coming Spr 18
- CSC 535 – Probabilistic Graphical Models
  - Kobus Barnard

2

Final Project

- Two options, choose **one**
  1. Explore your own model and data
  2. Implement MH for vision inference problem
- In both cases, concise, clear technical write-up

- **Make decision by Friday, Nov 17**
  – Email description to me

SISTA 3

Main parametric modeling frameworks

- Minimizing a Loss function
  – Linear model
  – Linear least mean squares
- Maximum Likelihood
  – Probabilistic model of uncertainty (noise, error)
  – Maximize the likelihood w.r.t. parameters
  – Linear model with additive Gaussian noise
- Bayesian Approach
  – Treat parameters as random variables
  – Use Bayes Theorem to combine likelihood & prior to find posterior distribution

Approaches to Model Selection
i.e., how to max ***generalization*** performance
Regularization (controlling model complexity)
Cross Validation (estimating the gen error)
Marginal Likelihood model selection

- Estimation Techniques (often used in Bayesian approaches)
  – Gradient methods (Widrow-Hoff (1st), Newton-Rhapson (2nd))
  – Laplace Approximation (estimate posterior with Gaussian)
  – Monte Carlo estimation of expectation; Metropolis-Hastings
- Classification (& Regression)
- Clustering
- Projection

predicting output

Main algorithmic families of Machine Learning

SISTA 4

Classification

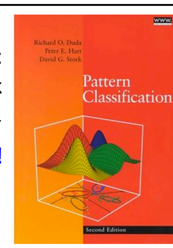
- N training objects, $\mathbf{x}_1, \ldots, \mathbf{x}_N$
- Each $\mathbf{x}_i$ is a D-dimensional vector
- Each object has a label, t_n describing the class object n belongs to
  - Typically class label is expressed as an integer
  - Binary case:
    - $t_n = \{0,1\}$ (logistic regression)
    - $t_n = \{-1,1\}$ (support vector machines)
  - C classes:
    - $t_n = \{1, 2, \ldots, C\}$ or $\{\{1,0,0,0\}, \{0,1,0,0\}, \{0,0,1,0\}, \ldots\}$
- Task: predict the class t_{new} for an unseen object $\mathbf{x}_{new}$

5

Issues in Classification

Duda
Hart
Stork
2001

Ch 2 !



- Different domains have different problems.
  - Disease Diagnosis: How do we handle the uneven cost of making errors?
  - Text classification: How do we handle complex data objects like text?

- Two very general ML approaches to Classification:
  - Non-probabilistic – if all we care about is class assignment (often, just define decision boundary)
  - Probabilistic – permits measure of *confidence* in class assignment

6

Probabilistic Classifiers

$$P(T_{\text{new}} = c | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$$

$$0 \leq P(T_{\text{new}} = c | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) \leq 1$$

$$\sum_{c=1}^{C} P(T_{\text{new}} = c | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = 1$$

Note: assumes mutual exclusivity!

Disease classification:

$$P(T_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = 0.6$$

versus

$$P(T_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = 0.9$$

SISTA 7

The Bayes Classifier

- Given a set of training points from *C* classes

$$P(T_{\text{new}} = c | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) =$$

$$P(T_{\text{new}} = c | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = \frac{p(\mathbf{x}_{\text{new}} | T_{\text{new}} = c, \mathbf{X}, \mathbf{t}) P(T_{\text{new}} = c | \mathbf{X}, \mathbf{t})}{\sum_{c'=1}^{C} p(\mathbf{x}_{\text{new}} | T_{\text{new}} = c', \mathbf{X}, \mathbf{t}) P(T_{\text{new}} = c' | \mathbf{X}, \mathbf{t})}$$

$$p(\mathbf{x}_{\text{new}} | T_{\text{new}} = c, \mathbf{X}, \mathbf{t}) \text{ and } P(T_{\text{new}} = c | \mathbf{X}, \mathbf{t})$$

Need to define *C* class-conditional distributions
Usually these are the same type (but not necessarily)
Then find parameters (MLE!)

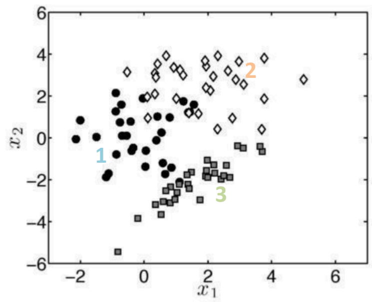
Probability of *c* “before evidence”
Can account for uneven class sizes (can bias for or against a class)
Always: must be positive and $\sum_c P(T_{\text{new}} = c | \mathbf{X}, \mathbf{t}) = 1$

1. Uniform prior: $P(T_{\text{new}} = c | \mathbf{X}, \mathbf{t}) = \frac{1}{C}$
2. Class size prior: $P(T_{\text{new}} = c | \mathbf{X}, \mathbf{t}) = \frac{N_c}{N}$

SISTA 8

The Bayes Classifier

- Example: **Gaussian** class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$

$$t_n = \{1, 2, 3\}$$

$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

One possibility: A Bayesian approach…

$$p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | \mathbf{X}^c) = \frac{p(\mathbf{X}^c | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{p(\mathbf{X}^c)}$$

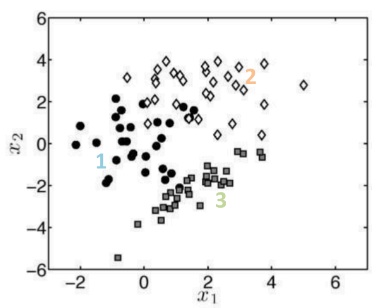
….then compute the likelihood of $\mathbf{x}_{new}$ by taking this expectation:

$$p(\mathbf{x}_{\text{new}} | T_{\text{new}} = c, \mathbf{X}, \mathbf{t}) = \mathbf{E}_{p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | \mathbf{X}^c)} \{ p(\mathbf{x}_{\text{new}} | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \}$$

This is most useful when there is little data and our estimates of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are uncertain

The Bayes Classifier

- Example: **Gaussian** class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$

$$t_n = \{1, 2, 3\}$$

$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

Another option: Direct maximum likelihood estimates of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} (\mathbf{x}_n - \boldsymbol{\mu}_c)(\mathbf{x}_n - \boldsymbol{\mu}_c)^\mathsf{T}$$

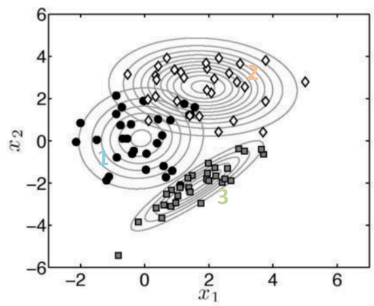
Summations are only for the data instances from the c^{th} class.

$$P(T_{\text{new}} = c | \mathbf{X}, \mathbf{t}) = \tfrac{1}{3}$$

SISTA

The Bayes Classifier

- Example: **Gaussian** class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$
$$t_n = \{1, 2, 3\}$$
$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

Another option: Direct maximum likelihood estimates of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} \mathbf{x}_n$$
$$\boldsymbol{\Sigma}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} (\mathbf{x}_n - \boldsymbol{\mu}_c)(\mathbf{x}_n - \boldsymbol{\mu}_c)^\mathsf{T}$$

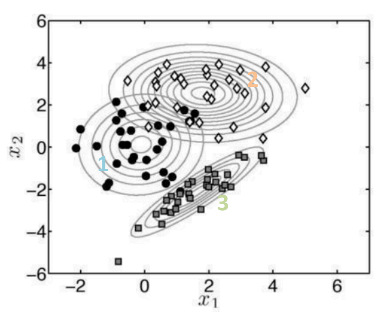
Summations are only for the data instances from the c^{th} class.

$$P(T_{\text{new}} = c | \mathbf{X}, \mathbf{t}) = \tfrac{1}{3}$$

SISTA 11

The Bayes Classifier

- Example: **Gaussian** class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$
$$t_n = \{1, 2, 3\}$$
$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

Making Predictions for $\mathbf{x}_{\text{new}} = [2, 0]^\mathsf{T}$

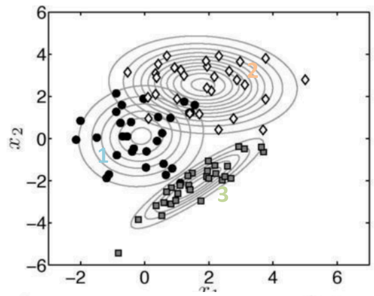
c	$p(\mathbf{x}_{\text{new}} \| T_{\text{new}} = c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$	$P(T_{\text{new}} = c \| \mathbf{X}, \mathbf{t})$	$p(\mathbf{x}_{\text{new}} \| T_{\text{new}} = c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) P(T_{\text{new}} = c \| \mathbf{X}, \mathbf{t})$		
1	0.0138	$\frac{1}{3}$	0.0046	0.6890	normalized
2	0.0061	$\frac{1}{3}$	0.0020	0.3024	
3	0.0002	$\frac{1}{3}$	0.0001	0.0087	

$$0.0046 + 0.0020 + 0.0001 = 0.0067$$

SISTA 12

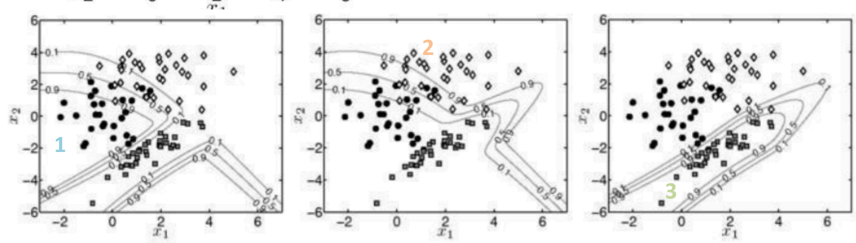
The Bayes Classifier

- Example: **Gaussian** class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$
$$t_n = \{1,\, 2,\, 3\}$$
$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$



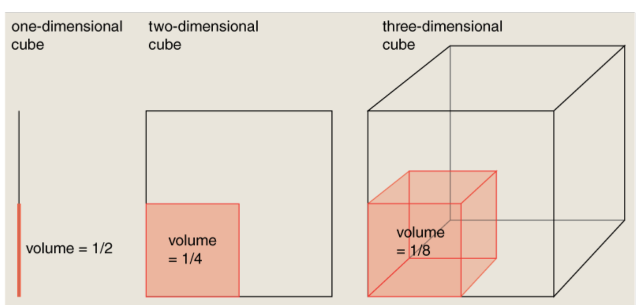
(a) $P(T_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$ (b) $P(T_{\text{new}} = 2 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$ (c) $P(T_{\text{new}} = 3 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$

https://erikbern.com/2015/10/20/nearest-neighbors-and-vector-models-epilogue-curse-of-dimensionality/

The Curse of Dimensionality

- Coined by Bellman, 1961.
- Actually refers to multiple phenomena found in a variety of domains: numerical analysis, sampling, combinatorics, data mining…
- **Common theme**: when the dimensionality increases, the volume of the space increases so fast that the available data become **sparse**.
- The amount of data needed generally grows exponentially with the dimensionality.

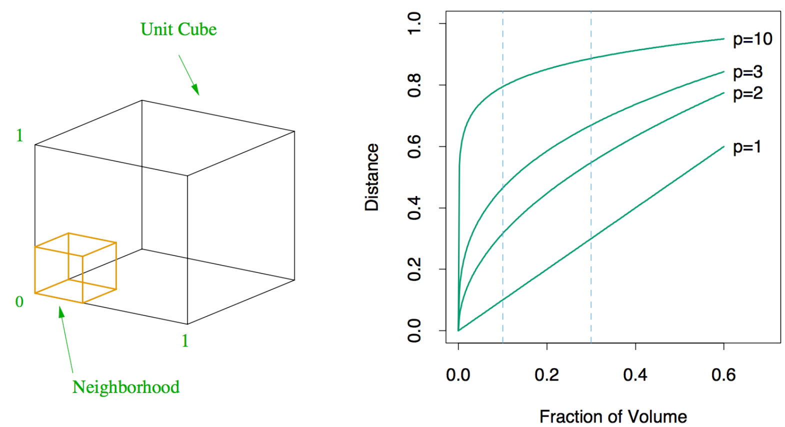
Some non-intuitive consequences of high dimensions



Consider all data within the "neighborhood" of a point (along each dimension, keep the "neighborhood" the same), and compare to total "volume".

The ratio of how much data is within that neighborhood decreases *rapidly* as the number of dimensions increases.

Some non-intuitive consequences of high dimensions

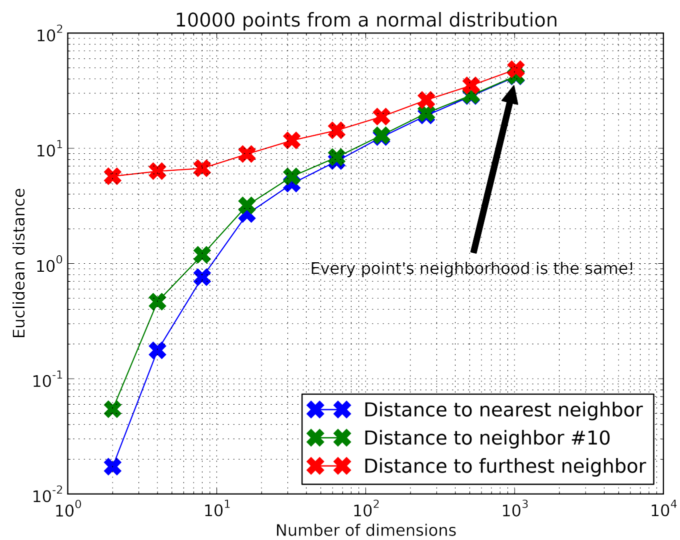


p = number of dimensions

In ten dimensions, need to cover 80% of the range of each coordinate to capture 10% of the data

17

Some non-intuitive consequences of high dimensions



18

The Bayes Classifier

- **Problem with Bayes Classifier**: Growth of parameters to estimate as number of dimensions (D) increases.
- For Gaussian class-conditional likelihood, modeling all covariance between features:

$$\underbrace{D}_{\text{For the mean}} + \underbrace{D + \frac{D(D-1)}{2}}_{\text{For the covariance matrix (diagonally symmetrical)}}$$

For 10 dimensions, 30 data points are not sufficient to estimate 65 parameters!

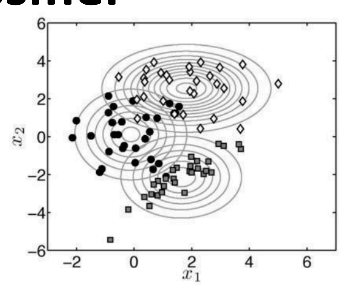
The key cost here is the covariance estimate.
We can dramatically simplify by assuming feature-conditional ***independence***

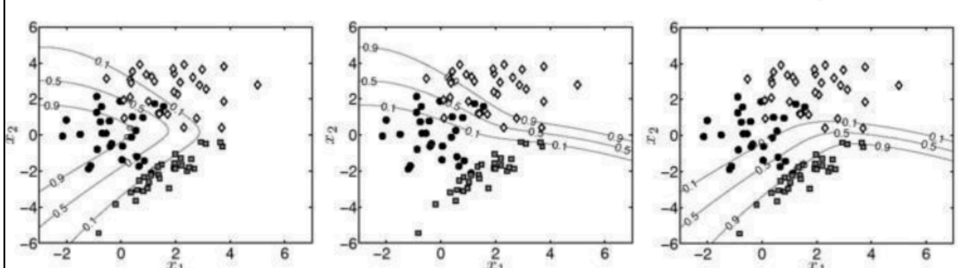
19

Naïve Bayes Classifier

- The Gaussian class- and ***feature*-conditional** likelihood

$$p(\mathbf{x}_n|t_n = k, \mathbf{X}, \mathbf{t}) = \prod_{d=1}^{2} p(x_{nd}|t_n = k, \mathbf{X}, \mathbf{t})$$





(a) $P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$ (b) $P(T_{\text{new}} = 2|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$ (c) $P(T_{\text{new}} = 3|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$