CS395T: Large Scale Data-Mining
Lecture 10
This lecture is given by Mr. Shailesh Kumar

Scribe by Bobby (Bo) Xiao

Learning problems

1) Unsupervised learning (e.g. Clustering, grouping portioning)
- Find structure in data.
- Given a set X = {, x1, x2 … xn }
- "minimize " a cost function:

  e.g. in K means, $\sum_{x \in X} \| x - m(c(x)) \|$

  e.g. in Graph portioning, Cut (V_1,V_2)

2) Supervised learning --- (Classification, Regressions)
- **"Learn" a mapping I (Input Space) ⟶ Ω (Class Labels),** e.g.
  For example:
  - Binary Images ⟶ { 'A', 'B', 'C'…'Z' }
  - Disease Symptoms ⟶ { Diseases } , Documenter Vectors ⟶ Yahoo category ,
  - Web pages ⟶ { relevant, non relevant}) (Web Crawler)
  - IQ Level ⟶ (Graduate students or / not)

- **Formal definition of learning**,
  Given a training set : X = {x_i , w(x_i) } i = 1 to n
  Learn a mapping : f : I ⟶ Ω s.t.
  For a new input x, f(x) is the class label assign to x
  So, it is the ability to generate to novel examples = proof of learning.

Two stage process:

1) Feature Selection/Extraction: I (Input space) → F (Feature Space)
   a) *Features selection*: Pick a small subset of inputs
      (e.g. removing stop words or word that occurs in all document)

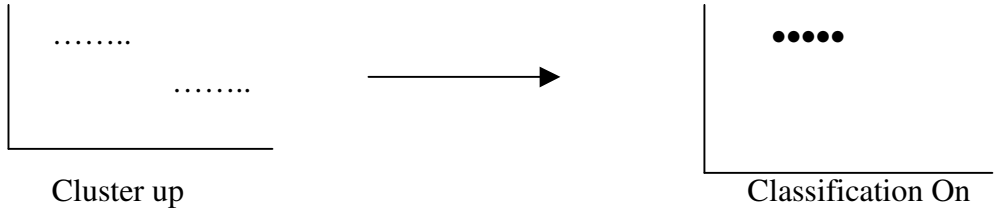
   b) *Features extraction*: Transform the inputs
      (e.g. Taking SVD and projecting on top K singular word, Fisher faces, Eigen faces)

Purpose:
1) Incorporate domain knowledge
2) Reduce input dimensionality from d.
3) Increase the speed of learning.

2) Classification: F (Feature Space) ⟶ Ω (Class labels)

Given a features vectors, assign a class label to it.

Different class label assigned to points



Simplest K NN Classifier:

--- Store all training data { xi, w(xi) } i = 1 … n

--- For any new input x, find its k nearest neighbors,

Let N (x) = { y1, y2, … yk } be k-NN's of x

Let w(N(x))) = {w(y1), w(y2),…,(yk)} be their class label

--- Count how many times each class label occurs in w(N(x)),

--- Let $n_c(x)$ = # of times class occurs in N(x) , then

$w(x) = \arg\max\ n_c(x),\ c = 1\ldots C$

Neighborhood: Can be defined in terms of evaluated distance or dot product similarity or any similarity measures features of K-NN classifiers

Characteristics of Nearest Neighbor classifier:

1) Stores all training data points (so no learning is involved)
2) For every new x, does a table look up : O(n)
3) Good for low dimensional space.
4) No parametric model is learned (typical examples of a non-parametric classifier)

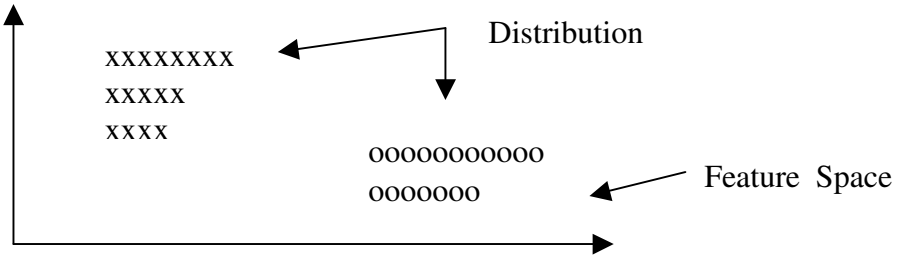
Humans:

1) We do not store all past examples to predict classify some examples.
2) We learn a model or concept for what a class is.

Model based (parametric) Classifiers

Probabilistic formulations (why?)

- No two inputs even of the same class can be the 'exactly' same.
- They "look" similar → have the same distributaries.
- {*A*, A, **A**, *A*, A } etc all in the same class.



- Points are scattered in a feature space.
- Points that belong in the same class are expected to be "closer" than points in different class
- Instead of saving all points, we could have a model for the "distribution " of points or a separating boundary

- e.g. 1) Learn the mean & covariance of each class.
2) Learn the plane (or more complex) boundary separating the two examples.

Data points: (Points not necessarily Elucidation always)
-- Discrete nominal (e.g. words, phrases, links, color)
-- Real values (e.g. temp, age, vector space model, frequency of occurrence)

Probabilistic Formulation of Classification Problems.

- C total # of class.
- Let ni = # of examples in class wi

N = n1 + n2 + n3 + … nc

X = X1∪ X2 ∪ ... ∪Xc are n examples, such that ni = |Xi|

Probabilities:

1) **Prior Probability**: of observing class wi: P(wi)
without seeing the data what can we say about any random point being in a class wi?

Examples:
Q: What is the probability that a random students is a CS grad students ? (P(CS-grad), P (other) }
Q: What is the probability that a random web page belongs to Yahoo Education or Sports?
Q: What is the probability that a random E-mail is a junk mail or a regular mail?
Q: P('A') > P('C') > P('Q') > …

Estimating prior probabilities:

$$P(wi)\frac{ni}{N} \rightarrow best_randomn_guess$$

P(w1) + P(w2) + .. P(wc) = 1

Ω = { w1, w2, w3,…, wc } = Exhaustive set of class.

2) **Posteriors probability**: of a given x belonging to class wi
P(wi | x) : Conditional probability of wi given observation x

Goal of learning to classify is equivalent to model P(wi | x)

Examples:
Q: P (next letter is 'u' | previous letter is 'q')
Q: P(CS-grad | IQ > 150)
Q: P(Education | {"grades, school" })
Q: P('junk mail' | { 'Free', 'chance' 'win', ' $10000' })

Bayes Theory:

Let A (e.g. x) and B (e.g. w) be two events that are not independent of each other.

Then the dependence between them can be written in two ways:

P(x, w) = p(x |w)P(w) = P(w|x) p (x) .

So the Bayes Theorem is:

$$P(w \mid x) = \frac{p(x \mid w).p(w)}{p(x)}$$

Examples:

$$P(cs_grad \mid IQ > 150) = \frac{P(IQ > 150 \mid cs_grad).p(cs_grad)}{p(IQ > 150)}$$

p(x|w) = class conditional probability distribution function.

p is small when x is continuous variable,
p is capital when x is nominal variable,

p(x) = p(x|w1) . P(w1) + p(x|w2) P(w2) + p(x|w3) P(w3) + ... + p(x|wc)P(wc)

Yahoo example:

	a1 = grade ,	a2 = score	
ω 1	2424	112	Education (Ne)
ω 2	145	842	Sports (Ns)
	572	74	Others (No)
Total	3141	Total 1028	N = Ne + Ns + No

$$P(a1 = \text{grade}) = \frac{3141}{N} \qquad P(a2 = \text{score}) = \frac{1028}{N}$$

$$P(\text{grade} \mid \text{"education"}) = \frac{2424}{Ne}$$

$$P(\text{score} \mid \text{"education"}) = \frac{112}{Ne}$$

$$P(\text{grade} \mid \text{"sports"}) = \frac{145}{Ns}$$

$$P(\text{score} \mid \text{"sports"}) = \frac{842}{Ns}$$

$$P(\text{education}\,|\,\{\text{grades}\}) = \frac{P(\text{grade}\,|\,\text{education}).P(\text{education})}{P(\text{grade})} = \frac{\frac{2424}{Ne}.\frac{Ne}{N}}{3141/N} = \frac{2424}{3141}$$

Using Bayes Rules for Class Fractions:

If for any x we know
P (w1 | x) = 0.7
P (w2 | x) = 0.25
P (w3 | x) = 0.05
Then , w(x) = w1 = arg max P (wi| x)
It is called **Maximum Aposteriori Probability Rule (MAP rule)**

What do we need to learn:
1) P(wi) simple
2) p(x|wi) for i = 1… C focus of various Supervised learning algorithm.

x = vector of d attributes < a1, a2, a3… ad >
e.g.
ai = ith word in a document
ai = ith line of a word page
ai = ith symptom of a disease.
ai = ith pix on an images (face recognition formulation)

ai ⟶ discrete ⟶ Machine learning. ⟶ {Naive Bayes Classifiers Bayesian networks}

ai ⟶ continuous ⟶ Statistical Pattern recognition. ⟶ {linear discriminant analysis Maximum likelihood classifiers.}

Naive Bayes Classifiers

The Naïve Bayes classifier assumes that all attributes are independent of each other conditioned on the class labels.

Naive assumption:

$$P(a1, a2, ad\,|\,wi) = P(wi)P(a1|wi)\,P(a2|wi)P(a3|wi)\ldots P(ad|wi)$$
$$= P(wi)\prod_{j=1}^{d} P(aj\,|\,wi)$$

i.e. given the class, all the attributes are independent of each other.

How do we compute P(aj | wi)

jth attributes can take (say) 2 values aj

X1		1	
X2		0	
.		0	
..		1	
		..	
Xn		1	

10 0's
90 1's

Simple counting down the jth column in x, is needed to compute P(aj | wi)

Xi points of class i: P (aj = 0 |wi) = 0.1 , p (aj= 1| wi) = 0.9

Example:

$$P(\text{education}\,|\,\{\text{grades},\text{scores}\})$$
$$= P(\text{education}).[P(\text{grade}\,|\,\text{education}).P(\text{scores}\,|\,\text{education})]$$
$$= \frac{\frac{Ne}{N}\frac{2424}{Ne}\frac{112}{Ne}}{P(\text{grade})P(\text{score})} = \frac{2424\times 112}{N\frac{3141}{N}\frac{1024}{N}}$$

So for a new sample:

1) First compute p(aj|wi) for a new sample x,
   i.e. if xj = Vjk then, p(aj = Vjk |wi)
   j = 1…d i = 1…c
2) Compute $P(x\,|\,wi) = \prod_{j=1}^{d}(p(aj\,|\,wi)$ (to every i = 1…c)
3) Compute posteriors:
   p(wi | x) ∝ P(wi) p(x|wi) . (to every i = 1…c)
4) Use map rule w(x) = arg max . P(wi |x) (to every i = 1…c)

Model: Instead of storing all the data points, for a Naïve Bayes classifier, we need to save:

{P(aj = Vjk |wi) , P (wi)}, for all j = 1.. d , k = 1… mj, i = 1…c)
where mj = number of possible values that attributes aj can take.

Model space complexity: $O((\sum_{j=1}^{d} mj)C + C)$