CS 383 – Machine Learning

Probabilistic/Statistical Classification

Slides adapted from material created by E. Alpaydin
Prof. Mordohai, Prof. Greenstadt, Pattern Classification (2nd Ed.),
Pattern Recognition and Machine Learning

Objectives

- Inference
- Bayesian Learning
- Naïve Bayes

Statistical Learning

Statistical Learning

- For our first approaches to classification we will look at the probability and statistics of our labeled data to make predictions on new data
- Review the Prob/Stats Week 0 slides!
- Hopefully these methods matches some of your intuition about data

Inference

Inference

- Our statistical learning will start with the concept of *inference*
  - Given distribution of seen data, what can we *infer* about new data?
- Given evidence/features $x = [x_1, x_2, \ldots, x_D]$ what is the *likelihood* that our object came from class i?
  - This is written as:
    $P(y = i|feature_1 = x_1, feature_2 = x_2, \ldots, feature_D = x_D)$
  - We'll just abbreviate this as:
    $P(y = i|f_1 = x_1, f_2 = x_2, \ldots, f_D = x_D) = P(y = i|f = x)$
- We call this value $P(y = i|f = x)$ that we're trying to compute, the *posterior*

Inference

$$P(y = i | f_1 = x_1, f_2 = x_2, \dots, f_D = x_D)$$

- Recall from probability that this is a *conditional probability:*

  "Given the first feature has value x_1, the second has x_2, etc.. what is the probability that our class was i?"

- Also recall that $P(a, b, c) = P(a \land b \land c)$ is called the *joint probability.*

Inference

- From the rules of probability

$$P(y|x) = \frac{P(y \wedge x)}{P(x)} = \frac{P(y, x)}{P(x)}$$

  - Here we call $P(x)$ the *evidence*.
- So we can solve this inference problem as:

$$P(y = i | f_1 = x_1, .., f_D = x_D) = \frac{P(y = i, f_1 = x_1, \ldots, f_D = x_D)}{P(f_1 = x_1, \ldots f_D = x_D)}$$

- Our final probability is defined purely in terms of the joints
  - And given enough data we be able get the joints easily directly from our data!

The Joint Distribution

- How to make a joint distribution:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables, then the table will have 2^M rows)
2. Count how many times in your data each combination occurs
3. Normalize those counts by the total data size in order to arrive a probabilities.

Note: The sum of joints must be equal to one

Learning a Joint Distribution

- To Build a JD (joint distribution) table for your attributes in which the probabilities are unspecified just fill in each row with

$$P(row) = \frac{records\ matching\ row}{total\ number\ of\ records}$$

A	B	C	Prob
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Example

- This JD was obtained by learning from three attributes in the UCI "Adult" Census database

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Using the Joint

- Once you have the JD you can easily compute the probability of any logical expression involving your attributes
- Using the *law of total probability* we can compute $P(Y)$ as the sum of all probabilities jointed with Y:

$$P(Y) = \sum_i P(Y \cap x_i) = \sum_{rows\ with\ Y} P(row)$$

- Examples:
  - What is P(Poor)?
  - What is P(Poor Male)?

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Inference with the Joint

- As mentioned, now we can also easily compute joint/conditional probabilities using our definition of the joint:

$$P(y|x) = \frac{P(y \wedge x)}{P(x)} = \frac{\sum_{rows\ with\ y\ and\ x} P(row)}{\sum_{rows\ with\ x} P(row)}$$

- What is $P(Male|Poor)$?

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Example

- Suppose we want to figure out given gender and hours worked, what is the wealth?
  - We can also write this as $P(W|G,H)$
- What is $P(W = rich|G = female, H = 40.5-)$?

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Inference is a big deal!

- There's tons of times you use it:
  - I've got this evidence. What's the chance that my conclusion is true?
  - I've got a sore neck. How likely am I to have Meningitis?
  - The lights are out and it's 9pm. What is the likelihood that my spouse is asleep?

Using Inference for Classification

- How can we use this for classification?
- Consider x, a set of D features
- To figure out which class a set of features should belong to we can just choose the class that maximizes the posterior probability

$$\hat{y} = \underset{i}{argmax}\, P(y = i | f = x)$$
$$= \underset{i}{argmax}\left(\frac{P(y = i, f = x)}{P(f = x)}\right)$$

Using Inference for Classification

$$\hat{y} = \underset{i}{argmax}\left(\frac{P(y = i, f = x)}{P(f = x)}\right)$$

- But since $P(f = x)$is the same for all classes we can just do:

$$\hat{y} = \underset{i}{argmax}\, P(y = i, f = x)$$

- And if we have $P(y = i, f = x)$ for all classes i then you can compute the actual probabilities, $P(y = i | f = x)$ by dividing by the sum of the joint probabilities:

$$P(y = i|x) = \frac{P(y = i, f = x)}{\sum_j P(y = j, f = x)}$$

Inference for Classification Example

- Given a rich male let's classify them as having worked more or less than 40.5 hours per week
  - $P(40.5+|male, rich) \propto P(40.5+, male, rich)$
  - $P(40.5-|male, rich) \propto P(40.5-, male, rich)$

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Bayesian Learning

Bayesian Decision Theory

- Sometimes we may be given information about how a class *generates* observations:
$$P(f = x|y = i)$$
- In these situations we can use Bayes' rule to solve $P(y = i|f = x)$ as
$$P(y = i|f = x) = \frac{P(y = i)P(f = x|y = i)}{P(f = x)}$$

Bayes Rule

$$P(y = i|f = x) = \frac{P(y = i)P(f = x|y = i)}{P(f = x)}$$

- In Bayes' Rule we call
  - $P(y = i|f = x)$ the posterior (what we want)
  - $P(y = i)$ the prior (probability that $y = i$)
  - $P(f = x|y = i)$ the likelihood (likelihood of generating x given y)
  - $P(f = x)$ the evidence

Bayesian Decision Theory

$$P(y = i|f = x) = \frac{P(y = i)P(f = x|y = i)}{P(f = x)}$$

- Prior $P(y = i)$ – comes from "prior" knowledge about the class, no data has been seen yet (it doesn't depend on x)
  - We have two classes C_1 and C_2 with probabilities $P(C_1)$, $P(C_2)$ such that $P(C_1) + P(C_2) = 1$
  - Therefore $P(y = 1) + P(y = 2) = 1$

Bayesian Decision Theory

$$P(y = i|f = x) = \frac{P(y = i)P(f = x|y = i)}{P(f = x)}$$

- Likelihood $P(f = x|y = i)$ - For class C_i, what's the "likelihood" of observing x
  - Or in other words, what's the likelihood that class C_i could generate x
  - Now we need observations to use!

Bayesian Decision Theory

$$P(y = i|f = x) = \frac{P(y = i)P(f = x|y = i)}{P(f = x)}$$

- Evidence $P(f = x)$ - How likely it is to observe x (regardless of class C_i)
  - If it's a rare observation it will increase our overall probability
  - Otherwise it's not that discriminatory, so it doesn't help much
- Posterior $P(y = i|f = x) = \frac{P(y=i)P(f = x|y = i)}{P(f=x)}$ tells us how likely to have class C_i given observation x
  - This is what we ultimately want!

Bayesian Classification

- Again for classification we want to choose
$$\hat{y} = \underset{i}{arg\max} P(y = i|f = x)$$
- And now with Bayesian Classification this becomes
$$\hat{y} = \underset{i}{arg\max} P(y = i)P(f = x|y = i)$$
  - Since the denominator $P(f = x)$ is a constant independent of the class
- And once again if we have $P(y = i)P(f = x|y = i)$ for all i we can compute the actual probabilities, $P(y = i|f = x)$ by dividing by the sum of the joint probabilities:
$$P(y = i|f = x) = \frac{P(y = i)P(f = x|y = i)}{\sum_j P(y = j)P(f = x|y = j)}$$

Bayesian Classifier Example

- Given Name, Height, Eye color, and Hair length let's try to guess (predict) the sex of a person

Name	Over 170CM	Eye	Hair length	Sex
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

Bayesian Classifier Example

- Ok so we want $P(S|N,T,E,H)$
- Suppose a person's name is Drew, is 160cm tall, has blue eyes and short hair
  - What's this person's sex?
- Compare
  - $P(S = male|Drew, <= 170cm, Blue, Short)$
  - $P(S = female|Drew, <= 170cm, Blue, Short)$
- Using inference:
  - $P(s|n,t,e,h) = \frac{P(s,n,t,e,h)}{P(n,t,e,h)}$

Name	Over 170CM	Eye	Hair length	Sex
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

Bayesian Classifier Example

- Let's do it using Bayes' Rule

$$P(s|n,t,e,h) = \frac{P(n,t,e,h|s)P(s)}{P(n,t,e,h)}$$

Name	Over 170CM	Eye	Hair length	Sex
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

Naïve Bayesian Inference/Classification

Conditional Independence

- Maybe it's tough to have enough data to reliably have either the joint $P(y = i, f = x)$ or the generative likelihood $P(f = x|y = i)$
- But if we make the assumption that all the features $X_{:1}, X_{:2}, \dots, X_{:D}$ are *conditionally* independent (perhaps **naively**), then this makes things a lot easier

Conditional Independence

- Two random variables, x and y are *conditionally independent* on z if given z, knowing y doesn't provide information about x
  - That's not to say they are necessarily individually independent of z, just that they their joint is conditionally independent of z
- If x and y are conditionally independent of z then we can say:

$$P(x, y|z) = P(x|z)P(y|z)$$

- Therefore

$$P(f = x|y = i) = \prod_{k=1}^{D} P(f_k = x_k|y = i)$$

Naïve Bayes Probability

- We can use the conditional probability to estimate the numerator of our Bayesian inference:

$$P(y=i|f=x)=\frac{P(y=i)P(x|y=i)}{P(f=x)}$$
$$=\frac{P(y=i)\prod_{k=1}^{D}P(f_k=x_k|y=i)}{P(f=x)}$$

- Remember, since the denominator is independent of the class, for classification we can just compute

$$P(y=i|f=x)\propto P(y=i)\prod_{k=1}^{D}P(f_k=x_k|y=i)$$

- Then dividing by the sum of these over all the classes

$$P(y=i|f=x)=\frac{P(y=i)\prod_{k=1}^{D}P(f_k=x_k|y=i)}{\sum_j P(y=j)\prod_{k=1}^{D}P(f_k=x_k|y=j)}$$

Example

- Let's try to determine if an object will be stolen
  - In particular $P(yes|red, sports, domestic)$
- Let's try it various different ways:
  - Infer $P(yes|red, sports, domestic)$
  - Bayesian Infer

    $P(yes|Color, Type, Origin)$
    $= P(Color, Type, Origin|Yes)P(Yes)/P(Color, Type, Origin)$
  - Naïve Bayes

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Classifying By Inference

- Which should we use?
- Depends on what you have
  - Ideally use regular inference
  - But if we don’t have complete joint information we may be able to use Bayesian Inference
    - Sometimes we’ll get the distributions that are created by a class.
  - If we have even less joint information and we can make an independence assumption then we can try naïve inference

Continuous Valued Data

Categorical vs Continuous Valued Data

- Each feature can typically fall into one of two categories:
    1. Categorical/finite-discrete – The feature can have one of M possible values (categories, enumerations)
    2. Continuous – The features can have any value!
- In all our inference work we had to essentially count how many times something occurred in order to compute its probability
    - This is only feasible for categorical data
- What if our data is continuous?

What if we have continuous χ_j?

- The general form of Naïve Bayes is:

$$P(y=i|f=x)=\frac{P(y=i)\prod_{j=1}^{D}P(f_k=x_k|y=i)}{\prod_{k=1}^{D}P(f_k=x_k)}$$

- But to do classification we might only need

$$P\,(y=i)\prod_{k}P(f_k=x_k|y=i)$$

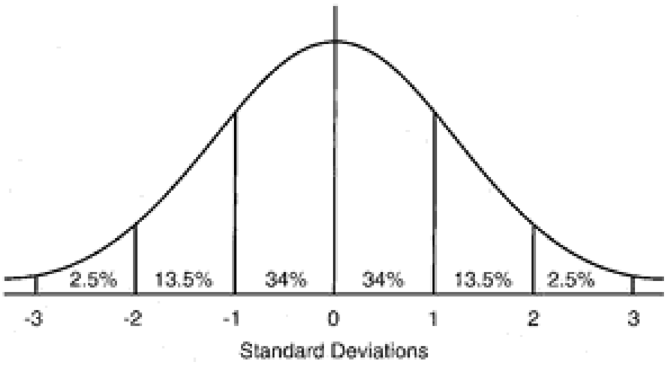
- How can we get $P(f_k=x_k|y=i)$ for x_k being continuous valued?
- Idea: Assume $P(f_k=x_k|y=i)$ follows a Normal (Gaussian) distribution

Recall Gaussian Distribution

- Quick review of Gaussian Distributions…
- $p(x|\mu,\sigma)$ is the *probability density function (PDF)* whose integral (over x) is 1

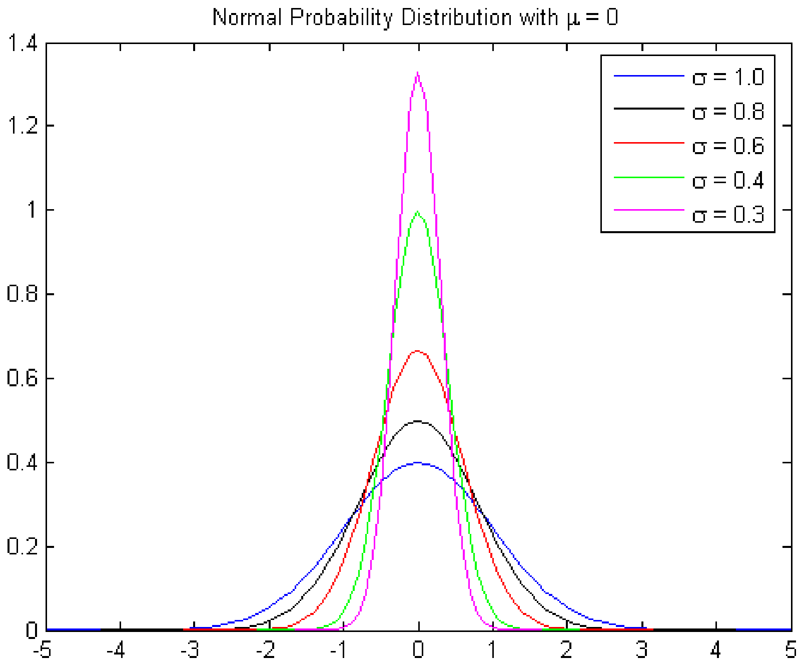
$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  - Where μ is the expected or mean value of x
  - Standard deviation is σ
- This value may be greater than 1, but should be proportional to the probability $P(x|\mu,\sigma)$.



Recall Gaussian Distribution

- Matlab:
  - `Mx = mean(X)`
  - `Sx = std(X)`
  - `px=normpdf(x,Mx,Sx)`



Gaussian Naïve Bayes: Continuous x, Discrete y

- How do we train a Naïve Bayes classifier with continuous features?
- For each discrete class C_i
  - First estimate the prior $P(y = i)$
  - For each attribute k, *estimate* $P(f_k = x_k | y = i)$ as $p(x_k | \mathcal{N}(\mu_{ik}, \sigma_{ik}))$by computing the attribute's mean and variance μ_{ik}, σ_{ik} from samples from that class C_i

Gaussian Naïve Bayes: Continuous x, Discrete y

- Now we can classify a new observation x as

$$\hat{y} = \underset{i}{argmax}\, P(y = i) \prod_k P(f_k = x_k | y = i)$$

$$\propto \underset{i}{argmax}\, P(y = i) \prod_k p(x_k | \mathcal{N}(\mu_{ik}, \sigma_{ik}))$$

- The product part of this, $\prod_k p(x_k | \mathcal{N}(\mu_{ik}, \sigma_{ik}))$, is sometimes referred to as the *maximum likelihood estimate (MLE)*.

Continuous Example

- Distinguish children from adults based on size
  - Classes $C = \{a, c\}$
  - Attributes: height[cm], weight[kg]
  - Training examples $\{h, w, y\}$, 4 adults, 12 children
- Class probabilities: $P(y = a)$?, $P(y = c)$?
- Model for adults:
  - Height ~Gaussian with
    - $\mu_{a,h} = \frac{1}{4}\sum_{i:y_i=a}(x_{i,h})$
    - $\sigma_{a,h}^2 = \frac{1}{4}\sum_{i:y_i=a}\left(x_{i,h} - \mu_{a,h}\right)^2$
  - Weight ~ Gaussian $\mathcal{N}(\mu_{a,w}, \sigma_{a,w})$
- Model for children…
  - Height ~ $\mathcal{N}(\mu_{c,h}, \sigma_{c,h})$
  - Weight ~ $\mathcal{N}(\mu_{c,w}, \sigma_{c,w})$

Continuous Example

- Now given a test sample $x = (w, h)$ we want to compute $P(y = a|f = x)$ and $P(y = c|f = x)$
- If we're using Bayes' approach then we're computing:
  - $P(y = a|f = x) = \frac{P(y=a)P(f = x|y = a)}{P(f=x)}$
  - $P(y = c|f = x) = \frac{P(y=c)P(f = x|y = c)}{P(f=x)}$
- Each has $P(f = x)$ in the denominator so let's ignore that
  - We can always compute the true probability by dividing by the sums if we need to.
- Then if we make a naïve independence assumption we arrive at
  - $P(y = a|f = x) \propto P(y = a)P(f_1 = w|y = a)P(f_2 = h|y = a)$
  - $P(y = c|f = x) \propto P(y = c)P(f_1 = w|y = c)P(f_2 = h|y = c)$

Continuous Example

- Approximations of $P(f_2 = h|y = a)$ as $p(f_2 = h|y = a)$, etc.. are from the PDFs of our Gaussians, etc..
- So finally
  - $P(y = a|f = x) \propto P(y = a)(p(h|\mathcal{N}(\mu_{a,h}, \sigma_{a,h}))p(w|\mathcal{N}(\mu_{a,w}, \sigma_{a,w})))$
  - $P(y = c|f = x) \propto P(y = c)(p(h|\mathcal{N}(\mu_{c,h}, \sigma_{c,h}))p(w|\mathcal{N}(\mu_{c,w}, \sigma_{c,w})))$
- If we want to threshold the probabilities then we likely want to normalize them so that $P(y = a|x) + P(y = c|x) = 1$
  - $P(y = a|f = x) = \frac{P(y = a|f = x)}{P(y = a|f = x) + P(y = b|f = x)}$
  - $P(y = c|f = x) = \frac{P(y = c|f = x)}{P(y = a|f = x) + P(y = b|f = x)}$

Multi-Class Classification

Multiple Classes

- Throughout our work with classification we'll typically focus on *binary* classification
  - Just two classes
    - (Positive, Negative), (0, 1), (1, 2), etc..
- For some applications this is fine.
- But in many applications we want to determine which of C $(C > 2)$ classes does an observation/instance belong to?
- Some classification approaches can handle this directly
  - For example in our probabilistic approaches seen here, we can just directly compute $P(y = i|f = x)$ for all classes $C_1, C_2, \ldots C_C$

Multiple Classes

- Other approaches only support binary classification directly.
- Fortunately there are two relatively straightforward ways of handling this:
  1. One vs All (need C classifiers) – often imbalanced, ambiguous regions
     - Choose the one that does best
  2. One vs one ($\frac{C(C-1)}{2}$ classifiers) – lots of classifiers
     - Fewer ambiguous regions
     - Choose the one that gets the most votes.
- We'll look at these approaches when they become necessary....

Final Observations

- Let’s think about this algorithm
  - Supervised or non?
  - Classification or regression?
  - Model-based or instance-based?
    - When it comes time to test/use, are we using the original data?
  - Linear vs Non-Linear?
  - Can this work on categorical data?
  - Can this work on continuous valued data?
  - Training Complexity?
  - Testing Complexity?
  - How to deal with overfitting?
  - Directly handles multi-class?