Naïve Bayes Classification

From the text book and from other lecture notes on Bayesian

1

Background

- Important ML taxonomy for learning models

probabilistic models vs non-probabilistic models

discriminative models vs generative models

2

2

Probability Basics

- Prior, conditional and joint probability for random variables
  - Prior probability: $P(x)$
  - Conditional probability: $P(x_1 | x_2), P(x_2 | x_1)$
  - Joint probability: $\mathbf{x} = (x_1, x_2),\ P(\mathbf{x}) = P(x_1, x_2)$
  - Relationship: $P(x_1, x_2) = P(x_2 | x_1)P(x_1) = P(x_1 | x_2)P(x_2)$
  - Independence:

$$P(x_2 | x_1) = P(x_2),\ P(x_1 | x_2) = P(x_1),\ P(x_1, x_2) = P(x_1)P(x_2)$$

- Bayesian Rule

$$P(c | \mathbf{x}) = \frac{P(\mathbf{x} | c)P(c)}{P(\mathbf{x})} \qquad Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Discriminative

Generative

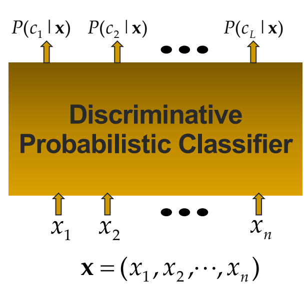
3

3

Probabilistic Classification Principle

- Establishing a probabilistic model for classification
  - **Discriminative model**

$$\boxed{P(c | \mathbf{x}) \quad c = c_1, \cdots, c_L, \mathbf{x} = (x_1, \cdots, x_n)}$$



- To train a discriminative classifier regardless its probabilistic or non-probabilistic nature**,** all training examples of different classes must be jointly used to build up a single discriminative classifier.
- Output L probabilities for L class labels in a probabilistic classifier while a single label is achieved by a non-

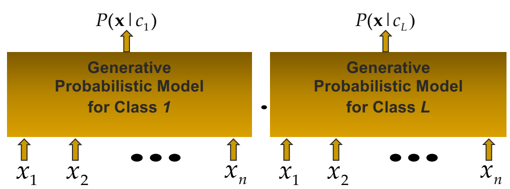
4

4

Probabilistic Classification Principle

- Establishing a probabilistic model for classification (cont.)
  - **Generative model (must be probabilistic)**

$$P(\mathbf{x} \mid c) \quad c = c_1, \cdots, c_L, \mathbf{x} = (x_1, \cdots, x_n)$$



$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

- L probabilistic models have to be trained independently
- Each is trained on only the examples of the same label
- Output L probabilities for a given input with L models
- "Generative" means that such a model produces data subject to the distribution via sampling.

5

5

Probabilistic Classification Principle

- **M**aximum **A** **P**osterior (**MAP**) classification rule
  - For an input x, find the largest one from L probabilities output by a discriminative probabilistic classifier $P(c_1 \mid \mathbf{x}), ..., P(c_L \mid \mathbf{x})$.
  - Assign x to label c^* if $P(c^* \mid \mathbf{x})$ is the largest.
- Generative classification with the MAP rule
  - Apply Bayesian rule to convert them into posterior probabilities

$$P(c_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_i)P(c_i)}{P(\mathbf{x})} \propto P(\mathbf{x} \mid c_i)P(c_i)$$

$$\text{for } i = 1, 2, \cdots, L$$

Common factor for all L probabilities

  - Then apply the MAP rule to assign a label

6

6

Naïve Bayes

- Bayes classification

$$P(c\,|\,\mathbf{x}) \propto P(\mathbf{x}\,|\,c)P(c) = P(x_1,\cdots,x_n\,|\,c)P(c) \ \text{ for } c = c_1,\ldots,c_L.$$

Difficulty: learning the joint probability $P(x_1,\cdots,x_n\,|\,c)$ is infeasible!

- Naïve Bayes classification
  - Assume all input features are class conditionally independent!

$$\begin{aligned} P(x_1,x_2,\cdots,x_n\,|\,c) &= P(x_1\,|\,x_2,\cdots,x_n,c)P(x_2,\cdots,x_n\,|\,c) \\ &= P(x_1\,|\,c)P(x_2,\cdots,x_n\,|\,c) \\ &= P(x_1\,|\,c)P(x_2\,|\,c)\cdots P(x_n\,|\,c) \end{aligned}$$

  - Apply the MAP classification rule: assign $\mathbf{x}' = (a_1,a_2,\cdots,a_n)$ to c^* if

$$[P(a_1\,|\,c^*)\cdots P(a_n\,|\,c^*)]P(c^*) > [P(a_1\,|\,c)\cdots P(a_n\,|\,c)]P(c), \quad c \neq c^*, c = c_1,\cdots,c_L$$

estimate of $P(a_1,\cdots,a_n\,|\,c^*)$

esitmate of $P(a_1,\cdots,a_n\,|\,c)$

7

7

Some Applications Include

- Identifying the types of iris flower
  - Given a set of features of iris flower, identify the specific class.
- Medical Diagnosis
  - Given a list of symptoms, predict whether a patient has cancer or not
- Weather
  - Based on temperature, humidity, etc… predict if it will rain tomorrow

8

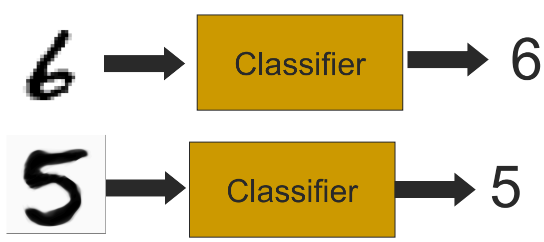
Bayesian Classification

- Problem statement:
  - Given features $X_1, X_2, \ldots, X_n$
  - Predict a label Y

9

Another Application

- Digit Recognition (from the MNIST fata set)



- $X_1, \ldots, X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)

10

The Bayes Classifier

- A good strategy is to predict:

$$\arg\max_{Y} P(Y|X_1, \ldots, X_n)$$

  - (for example: what is the probability that the image represents a 5 given its pixels?)

- So … how do we compute that?

11

The Bayes Classifier

- Use Bayes Rule!

$$P(Y|X_1, \ldots, X_n) = \frac{\overbrace{P(X_1, \ldots, X_n|Y)}^{\text{Likelihood}}\overbrace{P(Y)}^{\text{Prior}}}{\underbrace{P(X_1, \ldots, X_n)}_{\text{Normalization Constant}}}$$

- Why did this help? Well, we think that we might be able to specify how features are "generated" by the class label

12

The Bayes Classifier

- Let's expand this for our digit recognition task:

$$
\begin{aligned}
P(Y=5|X_1,\ldots,X_n) &= \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)} \\
P(Y=6|X_1,\ldots,X_n) &= \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}
\end{aligned}
$$

- To classify, we'll simply compute these two probabilities and predict based on which one is greater

13

Model Parameters

- For the Bayes classifier, we need to "learn" two functions, the likelihood and the prior

- How many parameters are required to specify the prior for our digit recognition example?

Just 1

14

Model Parameters

- How many parameters are required to specify the likelihood?
  - (Supposing that each image is 30x30 pixels)

$2(2^{900}-1)$

15

Model Parameters

- The problem with explicitly modeling $P(X_1,\ldots,X_n|Y)$ is that there are usually way too many parameters:
  - We'll run out of space
  - We'll run out of time
  - And we'll need tons of training data (which is usually not available)

16

The Naïve Bayes Model

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**
- Equationally speaking:

$$P(X_1, \ldots, X_n | Y) = \prod_{i=1}^{n} P(X_i | Y)$$

- (We will discuss the validity of this assumption later)

17

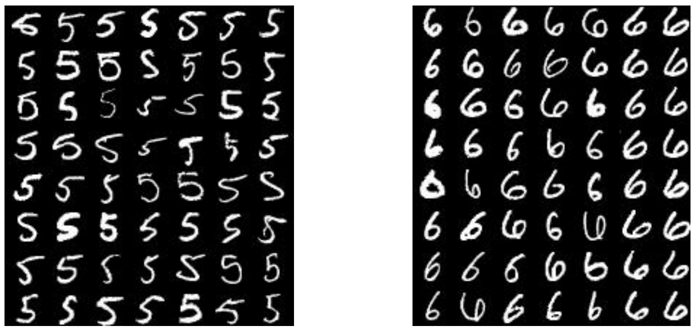
Why is this useful?

- # of parameters for modeling $P(X_1,\ldots,X_n|Y)$:
  - $2(2^n-1)$
- # of parameters for modeling $P(X_1|Y),\ldots,P(X_n|Y)$
  - $2n$

18

Naïve Bayes Training

- Now that we've decided to use a Naïve Bayes classifier, we need to train it with some data:



MNIST Training Data

Naïve Bayes Training

- Training in Naïve Bayes is **easy**:
  - Estimate P(Y=v) as the fraction of records with Y=v

$$P(Y = v) = \frac{Count(Y = v)}{\#\ records}$$

  - Estimate $P(X_i=u|Y=v)$ as the fraction of records with Y=v for which $X_i=u$

$$P(X_i = u|Y = v) = \frac{Count(X_i = u \wedge Y = v)}{Count(Y = v)}$$

- (This corresponds to Maximum Likelihood estimation of model parameters)

Naïve Bayes Training

- In practice, some of these counts can be zero
- Fix this by adding "virtual" counts:

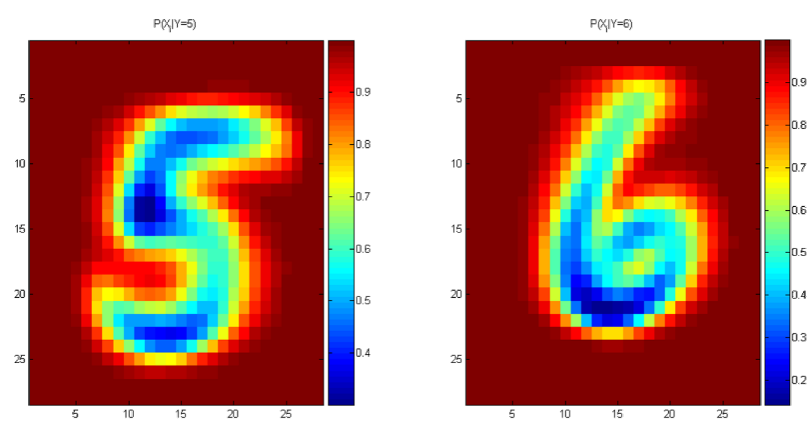
$$P(X_i = u | Y = v) = \frac{Count(X_i = u \wedge Y = v) + 1}{Count(Y = v) + 2}$$

  - (This is like putting a prior on parameters and doing MAP estimation instead of MLE)
  - This is called *Smoothing*

21

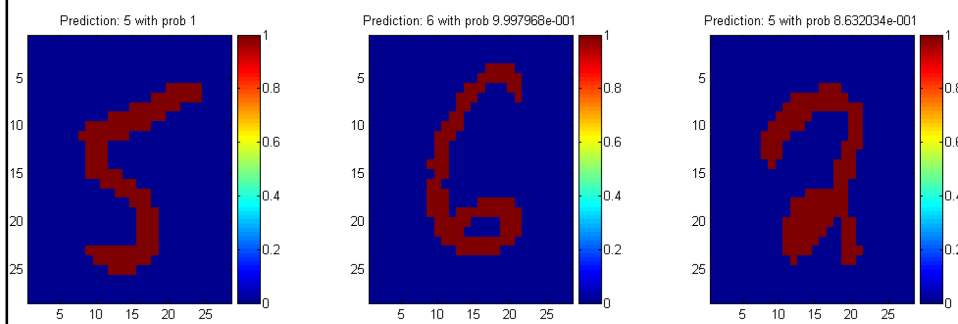
Naïve Bayes Training

- For binary digits, training amounts to averaging all of the training fives together and all of the training sixes together.



22

Naïve Bayes Classification



23

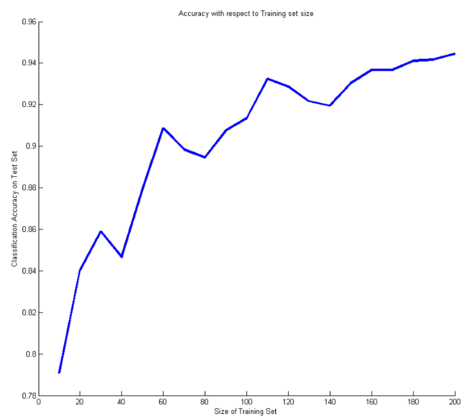
Outputting Probabilities

- What's nice about Naïve Bayes (and generative models in general) is that it returns probabilities
  - These probabilities can tell us how confident the algorithm is
  - So… don't throw away those probabilities!

24

Performance on a Test Set

- Naïve Bayes is often a good choice if you don't have much training data!



25

Naïve Bayes Assumption

- Recall the Naïve Bayes assumption:
  - that all features are independent **given the class label Y**
- Does this hold for the digit recognition problem?

26

Exclusive-OR Example

- For an example where conditional independence fails:
  - $Y = XOR(X_1, X_2)$

X_1	X_2	$P(Y=0\|X_1,X_2)$	$P(Y=1\|X_1,X_2)$
0	0	1	0
0	1	0	1
1	0	0	1
1	1	1	0

27

- Actually, the Naïve Bayes assumption is almost never true

- Still… Naïve Bayes often performs surprisingly well even when its assumptions do not hold

28

Numerical Stability

- It is often the case that machine learning algorithms need to work with very small numbers
  - Imagine computing the probability of 2000 independent coin flips
  - MATLAB thinks that $(.5)^{2000}=0$

29

Numerical Stability

- Instead of comparing $P(Y=5|X_1,\ldots,X_n)$ with $P(Y=6|X_1,\ldots,X_n)$,
  - Compare their logarithms

$$\begin{aligned}\log\left(P(Y|X_1,\ldots,X_n)\right) &= \log\left(\frac{P(X_1,\ldots,X_n|Y)\cdot P(Y)}{P(X_1,\ldots,X_n)}\right)\\ &= \text{constant} + \log\left(\prod_{i=1}^{n} P(X_i|Y)\right) + \log P(Y)\\ &= \text{constant} + \sum_{i=1}^{n}\log P(X_i|Y) + \log P(Y)\end{aligned}$$

30

Recovering the Probabilities

- Suppose that for some constant K, we have:

$$\log P(Y = 5 | X_1, \ldots, X_n) + K$$

  - And

$$\log P(Y = 6 | X_1, \ldots, X_n) + K$$

- How would we recover the original probabilities?

31

Recovering the Probabilities

- Given: $\alpha_i = \log a_i + K$
- Then for any constant C:

$$\begin{aligned} \frac{a_i}{\sum_i a_i} &= \frac{e^{\alpha_i}}{\sum_i e^{\alpha_i}} \\ &= \frac{e^C \cdot e^{\alpha_i}}{\sum_i e^C \cdot e^{\alpha_i}} \\ &= \frac{e^{\alpha_i + C}}{\sum_i e^{\alpha_i + C}} \end{aligned}$$

- One suggestion: set C such that the greatest α_i is shifted to zero:

$$C = -\max_i \{\alpha_i\}$$

32

Recap

- We defined a *Bayes classifier* but saw that it's intractable to compute $P(X_1,\ldots,X_n|Y)$
- We then used the *Naïve Bayes assumption* – that everything is independent given the class label Y
- A natural question: is there some happy compromise where we only assume that *some* features are conditionally independent?

33

Conclusions

- Naïve Bayes is:
  - Really easy to implement and often works well
  - Often a good first thing to try
  - Commonly used as a "punching bag" for smarter algorithms

34