ECE M146 Introduction to Machine Learning

Prof. Lara Dolecek

ECE Department, UCLA

Today’s Lecture

Recap:

- Classification methods so far

New topics:

- Generative modeling
- Naïve Bayes Classifier

Today’s Lecture

Recap:

- Classification methods so far

New topics:

- Generative modeling
- Naïve Bayes Classifier

Recap – modeling classification problems

- We have previously studied parametric methods with a discriminant function.
- What were these methods ?

- We have also studied a method with discriminative modeling.
- What was this method ?

Today's Lecture

Recap:

- Classification methods so far

New topics:

- Generative modeling
- Naïve Bayes Classifier

Alternative viewpoint

- Another way to model the inference system is to make the following assumption:

- There is some underlying joint distribution p(x,y) that jointly specifies x and y.
- Write it as:

- Recall the Bayes rule:

Generative modeling

- Advantage: can generate new data points (as many as you want) based on this model

- Disadvantage: once we make a choice what distribution is data from (e.g.,Bernoulli, Gaussian, etc.), that assumption stays forever.

Generative modeling

- Since we have a probabilistic modeling, we will typically look to maximize the following:

- This is known as maximizing the joint likelihood.
- When more convenient, we will maximize the log of the above expression.
- Maximization done by taking derivatives.

Today’s Lecture

Recap:

- Classification methods so far

New topics:

- Generative modeling
- Naïve Bayes Classifier

Naïve Bayes Classifier

- The key assumption is **conditional independence**.
- Math:

Practical example:

- Email spam filtering based on the words present in the email.
- Suppose x_1, x_2,…,x_d are indicators of d dictionary words, so that x_i is 1 if the i-th dictionary word is present, and x_i is 0 if absent.
- Naïve Bayes assumption:

A quick example of conditional independence

Likelihood

- Consider the likelihood for one example (e.g., one email):

- Now suppose we have N examples (e.g., N emails):

Derivations for N=1 example

- Suppose we are given prior probabilities for the label, as follows:

- Convenient representation of the above:

Derivations for N=1 example, continued

- Notation for conditional probabilities (4 total – why ?):

- What is the total number of parameters ?
- Answer: 2d+1

Derivations for N=1 example, continued

- Write the joint probability:

- Expand:

- Note that this is a compact representation of the joint probability: one of the two product terms will hold, but not both simultaneously.

Derivations for N=1 example, continued

- If we take the log, we get:

- This will also serve as an auxiliary result for what we will do next.

Let’s now consider the case of N examples

- Now, the likelihood is:
- Take the log so we get the double summation as follows:

Not as complicated as it looks!

- Useful notation:

Estimation of the parameter θ_0

- We are taking the derivative of the log likelihood to find the optimal choice.
- Notice that only the initial terms have θ_0 so only these terms matter when we take the derivative.

- So we can isolate:

Estimation of the parameter θ_0

- When we take the derivative, we get:

- This makes sense!

Estimation of the parameter $\theta_{1,1}$

- Let's now consider parameter $\theta_{1,1}$:

- Again, can isolate the terms in the log likelihood that matter for the derivative:

Estimation of the parameter $\theta_{1,1}$

- Useful notation:

- Taking the derivative yields:

Estimation of the remaining parameters

- Analogous to the previous case:

Classification rule at testing time

- At test time, we have the following rule:

Classification rule at testing time

- At test time, we have the following rule:

Description of the decision boundary

Description of the decision boundary, ctd

Description of the decision boundary, ctd

Laplace smoothing

- What if there was no instance of x_j in the training set in either class ?
- Recall math:

Laplace smoothing

- Force the estimates to not be strictly zero (or one).

- Example: