8001652 Introduction to Pattern Recognition. Lecture 7: Dimensionality, Computational Complexity, and Hidden Markov Models

Jussi Tohka

jussi.tohka@tut.fi

Institute of Signal Processing

Tampere University of Technology

Problems of dimensionality

- In practical multicategory applications, it is not at all unusual to encounter problems involving fifty or a hundred features.
- We might think that each feature is useful for at least some of the discriminations.
- There are two issues that must be confronted
  - How the classification accuracy depends upon the dimensionality (i.e. number of features) and amount of training data
  - Computational complexity of designing the classifier

Accuracy and Dimension

- If the features are independent, there are some theoretical results that suggest the possibility of excellent performance.
- As an example, consider the linear, two-class multivariate normal case.
- That is, the class conditional densities are Gaussian with equal covariance matrices:

$$p(\mathbf{x}|\omega_j) = N(\mu_j|\Sigma), j = 1, 2$$

- We further assume that the prior probabilities are equal and we apply the zero-one loss function.

Accuracy and Dimension

- Bayes risk (i.e. the error produced by the optimal classifier) is then

$$P(error) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{-u^2/2} du,$$

  where

$$r^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2).$$

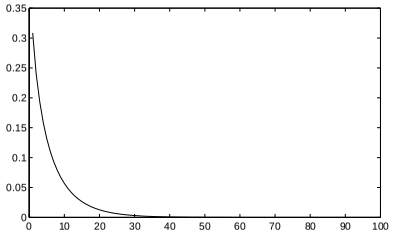
- r is the Mahalanobis distance between μ_1 and μ_2.
- $P(error)$ is related to cumulative distribution function of the normal distribution.

Accuracy and Dimension

- Let us assume that features are independent, that is the covariance matrix is diagonal.
- Then $r^2 = \sum_{i=1}^{d} \frac{(\mu_{1_i} - \mu_{2_i})^2}{\sigma_i^2}$.
- This shows how each (independent) feature reduces the classification error.
- Hence, in theory, if we just use many enough independent features, the Bayes risk can be made arbitrarily small.

Accuracy and Dimension

- Assume that the centers of the two classes are $\mu_1 = 0$ and $\mu_2 = 1$ and covariance matrix is an identity matrix.
- The Bayes risk is plotted as function of the number of features in a classifier below:



- With one feature, the Bayes risk is approx 31 % , with 100 features it is approx 0.00003 %.

Accuracy and Dimension

- Although in theory adding new features will increase the accuracy of the classifier, in practise it has been observed that beyond a certain point, additional features lead to worse performance.
- The basic source of the problem can be traced to the fact that we have a wrong model (e.g. Gaussian assumption) or the number of training samples is inadequate.
- However, precise analysis of the problem is very challenging.

Computational complexity

- It has been mentioned that one consideration affecting our design methodology is that of computational complexity.
- Therefore, the technical notion of computational complexity - so-called big-oh notation - can be useful.
- For further info on Big oh-notation look at www.cs.unc.edu/˜baruah/Teaching/2000f/Lectures/2000-09-05.pdf.
- Generally, classifier design is more computationally complex problem than classifier evaluation. In other words, learning the model for the class is more complex than deciding which model (or class) generated the measured features.

Hidden Markov Models

- So far, we have limited our attention to the problem of estimating the parameters in the class-conditional densities needed to make a single decision.
- In problems that have an inherent temporality, we may have states (of nature) that are influenced directly by previous the state of nature.
- HMMs (Hidden Markov Models) are tools for such problems.
- While the notation and mathematics are more complex, the underlying idea of classification remains the same.

Hidden Markov Models

- HMMs have a number of parameters whose values are set so as to best explain training patterns. This is similar to the task of estimating the parameters of class-conditional probability distributions based on the training data.
- Later, a test pattern is classified by the model that has the highest posterior probability, that is, that best explains the pattern.
- This is similar to the minimum error-rate classification.

First Order Markov Models

- We consider a sequence of states at successive times; the state at any time t is denoted $\omega(t)$.
- A sequence of states of length τ is $\omega^\tau = \{\omega(1), \omega(2), \ldots, \omega(\tau)\}$.
- Example: $\omega^3 = \{\omega_1, \omega_4, \omega_1\}$.
- The model for the production of any sequence is described by transition probabilities $P(\omega_j(t+1)|\omega_i(t)) = a_{ij}$. These time-independent probabilities describe the probability of transition from one state to another and fully specify a first-order Markov model (in case the initial state $\omega(0)$ is known).

First Order Markov Models

- Now assume that all a_{ij} (i.e. transition probabilities) are all known - that is we have a model θ.
- In first order Markov models we have $P(\omega(t+1)|\omega(t)) = P(\omega(t+1)|\omega(t), \ldots, \omega(1))$.
- We can then compute the probability that this model generated a particular sequence.
- In symbols $P(\omega^\tau|\theta) = \prod_{t=0}^{\tau-1} P(\omega(t+1)|\omega(t))$.
- Example: If $\omega^3 = \{\omega_1, \omega_4, \omega_1\}$ then the probability is $a_{14}a_{41}$.

First Order HMMs

- In addition to the standard Markov Model structure of hidden states (i.e. patterns to be recognized), HMMs emit a visible symbol at each instance of time.
- That is, we have two sequences $\mathbf{V}^\tau = \{v(1), v(2), \ldots, v(\tau)\}$ and ω^τ and also we know probabilities $P(v_k(t)|\omega^\tau) = P(v_k(t)|\omega_j(t)) = b_{jk}$.

The three problems

- **The evaluation problem**: Suppose we have an HMM, complete with transition probabilities a_{ij} and b_{jk}. Determine the probability that a particular sequence of visible states $\mathbf{V}^\tau$ was generated by that model.
- **The decoding problem**: Suppose we have an HMM as well as a set of observables $\mathbf{V}^\tau$. Determine the most likely hidden states that led to those observations.
- **The learning problem**: Suppose we are given the coarse structure of a model (the number of states and the number of visible states) but not the a_{ij}s and b_{jk}s. Give a set of training observations of visible symbols, determine these parameters.

The evaluation problem

- We shall consider only the evaluation problem, for other problems see the course book or a tutorial at www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.htm
- The probability that the model produces a sequence $\mathbf{V}^\tau$ of visible states is

$$P(\mathbf{V}^\tau) = \sum_r P(\mathbf{V}^\tau|\omega_r^\tau)P(\omega_r^\tau).$$

- Here the index r runs through all possible sequences of hidden states.

The evaluation problem

- But now $P(\omega^\tau) = \prod_{t=1}^{\tau} P(\omega(t)|\omega(t-1))$.
- And $P(\mathbf{V}^\tau|\omega^\tau) = \prod_{t=1}^{\tau} P(v(t)|\omega(t))$.
- Hence $P(\mathbf{V}^\tau) = \sum_r \prod_{t=1}^{\tau} P(\omega_r(t)|\omega_r(t-1))P(v(t)|\omega_r(t))$. of the above formula is not necessarily computationally tractable.
- However, there exist simple, fast algorithm for solving the evaluation problem (see the course book).