Phoneme Classification Using Support Vector Machine
Speech Communication (CT437) Project

Falak Shah (201311024)
Course Instructor: Prof. Hemant Patil

MTech, DA-IICT

November 19, 2014

Outline

Introduction and Objectives

- Phoneme classification using Support Vector Machine in framewise manner
- Application: frontend of an ASR system
- Approaches suggested for phoneme classification in [5],[2]
- Introduction to support vector machine

Introduction to SVM

- Arguably the most successful machine learning tool for classification.
- Optimization packages available for solving the classification problem of SVM.(LibSVM, SVMTorch)
- Intended application: classify 40 phoneme classes

Maximising the Margin

- Linearly Separable data- selecting the best margin out of possible margins
- Bigger margin better as even in case of noisy data crossover probability is less.

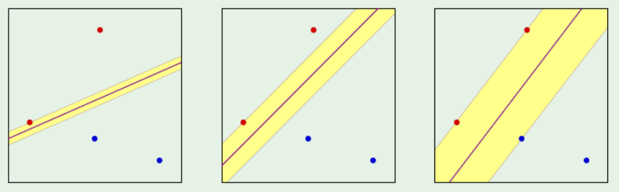

Figure 1: Bigger margin better

Finding w with large margin

- Task is to find w and b for the separating hyperplane $w^T x + b = 0$.
- Here, b denotes bias, $x \in R^d$ is the input feature vector then $w = \{w_1, w_2, w_3 ... w_d\}$.
- For linearly separable points, the plane will not touch any points, i.e. $|w^T x + b| > 0$.
- Scaling of w and b by same amount result in the same plane. So, we select w and b such that $|w^T x_n + b| = 1$, x_n being the point closest to the hyperplane.
- Here, we'll use euclidean distance as the yardstick for measurement.

Distance Computation

- Distance between the plane $w^T x + b = 0$ and the nearest point x_n is the margin. Given that, $|w^T x_n + b| = 1$.
- Result 1:Vector w is $\perp$ to the plane.
- Proof: For x' and x'' on the plane $w^T x' + b = 0$ and $w^T x'' + b = 0$. So, $w^T(x' - x'') = 0$

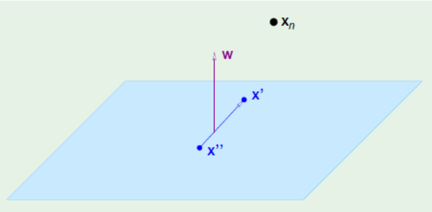

Figure 2: w $\perp$ to the plane

Distance Computation

- Projection of $x_n - x$ for any x on the plane onto the normal w.

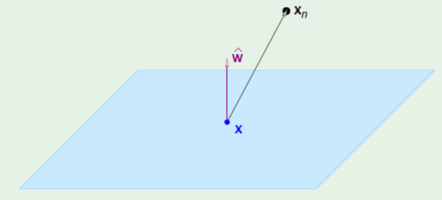

Figure 3: Dsitance from the plane

- Distance d is given by
  d$=\frac{1}{||w||}|w^T(x_n - x)|=\frac{1}{||w||}|(w^T x_n + b) - (w^T x + b)| = \frac{1}{||w||}$

Optimization problem

- Maximize $\frac{1}{||w||}$ subject to $\min_{n=1,2..N} |w^T x_n + b| = 1$.
- Also, $|w^T x_n + b| = y_n(w^T x_n + b)$, since we only consider the correctly classified data.
- In the alternative representation, we can write the problem as Minimize $\frac{1}{2}(w^T w)$ subject to $y_n(w^T x_n + b) >= 1$ for $n = 1, 2...N$
- This statement is equivalent to the above one as minimum value for will only be achieved when $y_n(w^T x_n + b) = 1$ because till then w and b can still be proportionately scaled down.

Constrained Optimization problem

- Minimize $\frac{1}{2}(w^T w)$ subject $y_n(w^T x_n + b) - 1 >= 0$
- Solution to this will yield the separating hyperplane with the largest margin.
- Constrained optimization problem converted to unconstrained optimization by lagrangian.
- KKT conditions needed for solution of lagrangian under inequality constraint.

Lagrangian and KKT conditions

- KKT approach generalizes the method of Lagrange multipliers, which allows only equality constraints.
- Given a problem as $\min_x f(x)$ subject to $g(x) <= 0$.
- Define the lagrangian as
  $L(x, \lambda) = f(x) + \lambda g(x)$ Then,
  x^* a local minimum $\iff$ there exists a unique λ^* s.t.
  1. $\nabla_x L(x^*, \lambda^*) = 0$
  2. $\lambda^* >= 0$
  3. $\lambda^* g(x^*) = 0$
  4. $g(x^*) <= 0$

Lagrangian formulation of the problem

- Minimize $\frac{1}{2}(w^T w)$ subject to $y_n(w^T x_n + b) - 1 >= 0$ for $n = 1, 2...N$
- Minimize $L(w, b, \alpha) = \frac{1}{2}(w^T w) - \sum_{n=1}^{N} \alpha_n(y_n(w^T x_n + b) - 1)$ w.r.t w and b and maximize w.r.t. each $\alpha_n >= 0$
- $\nabla_w L = w - \sum_{n=1}^{N} \alpha_n y_n x_n = 0$
- $\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0$
- KKT condition 3: $\alpha_n(y_n(w^T x_n + b) - 1) = 0$
- Substituting these values in the original equation results in the dual representation of the problem.

Lagrangian formulation of the problem

- Substituting $w = \sum_{n=1}^{N} \alpha_n y_n x_n$ and $\sum_{n=1}^{N} \alpha_n y_n = 0$ in the lagrangian
- $L(w, b, \alpha) = \frac{1}{2}(w^T w) - \sum_{n=1}^{N} \alpha_n (y_n (w^T x_n + b) - 1)$ we get,
- $L(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n^T x_m$
- Maximise w.r.t. α and subject to $\alpha_n >= 0$ for n=1,2,...N and $\sum_{n=1}^{N} \alpha_n y_n = 0$

Quadratic Programming

$$\max_{\alpha} L(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n^T x_m$$

Alternatively,

$$\min_{\alpha} L(\alpha) = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n^T x_m - \sum_{n=1}^{N} \alpha_n$$

subject to $\alpha_n >= 0$ for n=1,2,...N and $\sum_{n=1}^{N} \alpha_n y_n = 0$

Quadratic programming formulation

$$\min_{\alpha} \frac{1}{2}\alpha^T \begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \ldots & y_1 y_N x_1^T x_N \\ y_2 y_1 x_2^T x_1 & y_2 y_2 x_2^T x_2 & \ldots & y_2 y_N x_2^T x_N \\ \ldots. & \ldots. & \ldots.. & \ldots. \\ y_2 y_1 x_2^T x_1 & y_2 y_2 x_2^T x_2 & \ldots & y_2 y_N x_2^T x_N \end{bmatrix} \alpha + (-1)^T \alpha$$

subject to a linear constraint $y^T \alpha = 0$ and range of

$0 <= \alpha <= \infty$ and

The size of the matrix depends on the size of the training dataset.

QP hands back α

- $\alpha = (\alpha_1, \alpha_2, ..., \alpha_N)$
- $w = \sum_{n=1}^{N} \alpha_n y_n x_n$
- α is a sparse vector. Sice we've the KKT condition $\alpha_n(y_n(w^T x_n + b) - 1) = 0$.
- Either $\alpha_n = 0$ for the interior points or $y_n(w^T x_n + b) = 1$ for the suppport vectors- the only ones where $\alpha_n > 0$. The x_n for which $\alpha_n > 0$ are called the support vectors as they only contribute to the solution. So now,
- $w = \sum_{x_n \in S.V.} \alpha_n y_n x_n$
- Solve for any b using $y_n(w^T x_n + b) = 1$. This will give same b for any S.V. This is also verification that the task is correctly accomplished.

Linearly non-separable data

- Cover's theorem: The probability that classes are linearly separable increases when the features are nonlinearly mapped to a higher dimensional feature space. [3]
- Proof:For N samples in l-dimensional feature space, the number of dichotomies (linearly separable groupings) is [1]

$$O(N, l) = 2 \sum_{i=0}^{l} \binom{N-1}{i}$$

- The total number of groupings is 2^N. Thus, the probability that the samples are linearly separable is the ratio

$$P_N^l = \frac{O(N, l)}{2^N}$$

Kernel Function

- Given 2 points x and x', we need z and z'. Where, $z = \phi(x)$
- Let $z^T z = K(x, x')$- the kernel function
- The trick is computing $K(x, x')$ without transforming x and x'.
- The function K can be arbitrarily chosen as long as the existence of $\phi(.)$ is guarnateed.
- A kernel K:$X \times X \rightarrow R$ is positive definite symmetric.

Kernel Function

- Mercer's condition: There exists a mapping $\phi(.)$ if and only if, for any g(x) such that

$$\int g(x)^2 dx$$

is finite then

$$K(x,y)g(x)g(y)dxdy \geq 0.$$

- Any kernel which can be expressed as $K(x,y) = \sum_{p=0}^{\infty} c_p (x.y)^p$, where the c_p are positive real coefficients and the series is convergent, satisfies the condition
- RBF kernel: $K(x,x') = exp(-\gamma||x - x'||^2)$
- Infinite dimensional z: with $\gamma = 1$ $K(x,x') = exp(-x^2)exp(-x'^2) \sum_{k=0}^{\infty} \frac{2^k (x)^k (x')^k}{k!}$

Framewise Phone Classification

- In [5] Jesper et al. describe use of Support Vector Machines for phonetic classification on the TIMIT corpus.
- Results in [2] show SVM outperform mixture of Gaussian based phonetic classification
- Framewise phone classification accuracy- comparable to other stae of the art
- Intended application: Frontend of an ASR system. The major issues are:
    1. Choice of kernel and its parameters
    2. Building multi-class classifiers from inherently binary SVM's

Experiment

- Experiment on the TIMIT corpus
- Features: Mel-scale cepstral coefficients (MFCCs) using 25ms frames spaced at 10ms intervals
- Input pattern x_i consists of the current frame of 12 MFCCs and energy plus delta and acceleration coefficients, and two context frames on each side, making a total of $(13 + 13 + 13) \times 5 = 195$ components.
- Testing on subset showed the Gaussian kernel to be the best one
- Parameters for kernel found by using exhaustive grid search on γ and c

Multiclass SVM

- One v/s all and one v/s all approaches possible
- One v/s one in which one classifier trained for all possible combinations with k=40 phoneme classes
- $\frac{1}{2}k(k-1) \implies$ 780 classifiers
- Two schemes to use these classifiers
  1. One v/s one voting scheme -simple majority vote
  2. Directed acyclic graph (DAGSVM) scheme- greedy decision graph based algorithm.
     Only $k-1=39$ classifications required.
- DAGSVM shown to outperform One v/s one [5]

Future Work

- An experiment using a smaller training set by including training samples from all classes
- In [4], Paliwal et al. show the usefulness of phase information even when STFT is taken for short windows as in ASR applications.
- Results of inclusion of phase information as features in SVM training

Christopher J. C. Burges.
A tutorial on support vector machines for pattern recognition.
Data Min. Knowl. Discov., 2(2):121–167, June 1998.

P. Clarkson and P.J. Moreno.
On the use of support vector machines for phonetic classification.
In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 2, pages 585–588 vol.2, Mar 1999.

T.M. Cover.
Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition.
Electronic Computers, IEEE Transactions on, EC-14(3):326–334, June 1965.

Kuldip K. Paliwal and Leigh Alsteris.
Usefulness of phase spectrum in human speech perception.
In *Proc. Eurospeech*, pages 2117–2120, 2003.

Jesper Salomon, Simon King, and Jesper Salomon.
Framewise phone classification using support vector machines.
In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, 2002.