KMeans Kernel Classifier

Course: Math Behind ML

Karthik Kurugodu
B.Tech Mathematics and Computing
Indian Institute of Technology Hyderabad
ma20btech11008@iith.ac.in

Nikhil Kongara
B.Tech Mathematics and Computing
Indian Institute of Technology Hyderabad
ma20btech11011@iith.ac.in

***Abstract*—The least squares SVM is a kernel method for non-linear regression and classification tasks. Here we combine KMeans clustering with the least squares SVM. First KMeans clustering is used to extract a set of representative vectors for each class, and then LS-SVM uses these representative vectors as a training dataset for the classification task**

I. Introduction

The kernel methods transform a given non-linear problem into a linear one by using a similarity kernel function $\Omega(x, x\prime)$. It is a similarity function defined over pairs of input data points $(x, x\prime)$. This way the input data is mapped into a higher dimensional feature space $\phi(x)$, where the inner product $\langle \cdot \ , \ \cdot \rangle$ can be calculated using Mercer's condition:

$$\Omega(x, x\prime) = \langle x \ , \ x\prime \rangle \tag{1}$$

Consider $\chi = \{x_n | n = 1, \ldots, N\}$ as training dataset.
Representer theorem: Any non-linear function $f : \chi \longrightarrow \mathbb{R}$ can be expressed as linear combination of kernel products on training dataset which was mentioned above earlier.

$$f(x) = \sum_{n=1}^{N} a_n \Omega(x, x_n) \tag{2}$$

Time complexity of LS-SVM is O(N^3) where N is size of the training dataset which is too high and makes it unsuitable for large dataset. So for this reason we use KMeans clustering to extract a set of representative vectors for each class, and then LS-SVM uses these representative vectors as a training dataset for the classification task. This way we can reduce the time complexity of LS-SVM to O(K^3) where K is the number of clusters. These representative vectors are called as **centroids**. These are then used by LS-SVM to classify the test data.
This KMeans-LS-SVM method has some advantages:

- It is faster than LS-SVM.
- It is more robust.
- It is very easy to implement.

II. Kernel LS-SVM Classifer

We already know that in binary classification, kernel SVM method constructs an hyperplane with the maximal margin between the two classes in feature space $\phi(x)$. This can be represented as convex quadratic programming problem involving inequality constraints.
The kernel LS-SVM simplifies the optimization problem by considering equality constraints only, such that solution is obtained by solving a system of linear equations. Now this problem is similar to ridge regression problem which is formulated as follows:

$$\min_{w,b} \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{n=1}^{N} (\hat{y}_n - w^T \phi(x_n) - b)^2 \tag{3}$$

Assume that K classes are encoded using standard basis in $\mathbb{R}^K$, i.e, let $x_i \in C_k$, then output y_i is a vector with 1 in the k^{th} position and 0 elsewhere:

$$y_{ij} = \begin{cases} 1 & \text{if } x_i \in C_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Consider input data $\{(x_i, yi) | x_i \in \mathbb{R}^{\mathbb{M}}, y_i \in \mathbb{R}^{\mathbb{K}}, i = 1, \ldots, N\}$ and the feature mapping function $\phi(x)$. The kernel LS-SVM is formulated as follows:

$$\min_{w,b} S(w, b, \epsilon) = \frac{1}{2} \sum_{j=1}^{K} w_j^T w_j + \frac{\gamma}{2} \sum_{i=1}^{N} \sum_{j=1}^{K} (\epsilon_{ij})^2 \tag{5}$$

subject to

$$\langle \phi(x)\ ,\ \omega_j \rangle + b_j = y_{ij} - \epsilon_{ij}, i = 1, \ldots, N; j = 1, \ldots, K \tag{6}$$

$$w_j^T \phi(x_i) + b_j = y_{ij} - \epsilon_{ij}, i = 1, \ldots, N; j = 1, \ldots, K \tag{7}$$

where $\epsilon_{ij} \geq 0$ are approximation errors, b_j is bias coefficient, $w^{(j)}$ is the vector of weights corresponding to the j^{th} class. The objective function S is a sum of least squares errors and the regularization term. This regularization parameter γ corresponds to a multi-dimensional version of the ridge regression problem.

In the primal weight space the multi class classifier takes the form:

$$x \in C_k, \Leftrightarrow k = arg \max_{j=1,\ldots,K} g_j(x) \tag{8}$$

$$\text{where } g_j(x) = \frac{\exp(\langle \phi(x)\ ,\ w^{(j)} \rangle + b_j)}{\sum_{i=1}^{K} \exp(\langle \phi(x)\ ,\ w^{(i)} \rangle + b_i)} \tag{9}$$

Here g_j is the non-linear soft max function

Now applying Lagrangian to (5)

$$L(w, b, \epsilon, a) = S(w, b, \epsilon) \\ - \sum_{i=1}^{N} \sum_{j=1}^{K} a_{ij} [\langle \phi(x)\ ,\ \omega_j \rangle + b_j - y_{ij} + \epsilon_{ij}]$$

where $a_{ij} \in \mathbb{R}$ is the lagrange multiplier. Now applying KKT conditions:

$$\frac{\partial L}{\partial w^{(j)}} = 0 \implies w^{(j)} = \sum_{n=1}^{N} a_{nj} \phi(x_n) \tag{10}$$

$$\frac{\partial L}{\partial b_{(j)}} = 0 \implies \sum_{i=1}^{N} a_{ij} = 0 \tag{11}$$

$$\frac{\partial L}{\partial \epsilon_{(ij)}} = 0 \implies a_{ij} = \gamma \epsilon_{ij} \tag{12}$$

$$\frac{\partial L}{\partial a_{(ij)}} = 0 \implies \langle \phi(x)\ ,\ \omega_j \rangle + b_j - y_{ij} + \epsilon_{ij} = 0 \tag{13}$$

Now from eq(10) , eq(12) and eq(13):

$$\sum_{n=1}^{N} [\Omega(x_i, x_n) + \gamma^{-1} \delta_{in}] a_{nj} + bj = y_{ij}, \tag{14}$$

Here δ_{in} is the Kronecker delta function: where $\delta_{in} = 1$ if $i = n$ and 0 otherwise

As you can see in eq(14) there are K independent system of equations with binary labels y_{ij}. Now each system can be written in the matrix form as follows:

$$\begin{bmatrix} 0 & u^T \\ u & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b_j \\ a^{(j)} \end{bmatrix} = \begin{bmatrix} 0 \\ y_j \end{bmatrix}, j = 1, \ldots, K \tag{15}$$

Here $I_{N \times N}$ is the identity matrix, $u_{N \times 1} = [1, \ldots, 1]^T$ is a vector of ones, $a_{N \times 1}^{(j)} = [a_{1j}, \ldots, a_{Nj}]^T$ is weights and $y_j = [y_{1j}, \ldots, y_{Nj}]^T$ is the vector of binary labels for the j^{th} class.

Each system has $N + 1$ linear equations with $N + 1$ unknowns. Each system has $N+1$ linear equations with $N+1$ unknowns.

$$\Theta = \begin{bmatrix} 0 & u^T \\ u & \Omega + \gamma^{-1} I \end{bmatrix} \tag{16}$$

All the K systems can be written as:

$$\Theta W = Z \tag{17}$$

where

$$W_{(N+1) \times K} = \begin{bmatrix} b_1 & \ldots & b_K \\ a^{(1)} & \ldots & a^{(K)} \end{bmatrix}, Z_{(N+1) \times K} = \begin{bmatrix} 0 & \ldots & 0 \\ y_1 & \ldots & y_K \end{bmatrix}$$

Now once all the K systems are solved, we consider multi-class classifier in dual space(from eq (14)) as follows:

$$g_j(x) = \frac{\exp(\langle \phi(x)\ ,\ w^{(j)} \rangle + b_j)}{\sum_{i=1}^{K} \exp(\langle \phi(x)\ ,\ w^{(i)} \rangle + b_i)}$$

From eq(9) and eq(10), we get:

$$g_j(x) = \frac{\sum_{n=1}^{N} \exp(\Omega(x, x_n) a_{nj} + b_j)}{\sum_{i=1}^{K} \sum_{n=1}^{N} \exp(\Omega(x, x_n) a_{ni} + b_i)}$$

Now our problem becomes:

$$x \in C_k, \Leftrightarrow k = arg \max_{j=1,\ldots,K} g_j(x) \tag{18}$$

where

$$g_j(x) = \frac{\sum_{n=1}^{N} \exp(\Omega(x, x_n)a_{nj} + b_j)}{\sum_{i=1}^{K}\sum_{n=1}^{N} \exp(\Omega(x, x_n)a_{ni} + b_i)}$$

Here g_j is the non-linear soft max function

III. KMeans Clustering

First we use KMeans clustering algorithm to extract a set of representative vectors for each class. Now this representative vectors will be passed into LS-SVM kernel model as training dataset. KMeans clustering algorithm is as follows:

1) Take $\{x_i^k | x_i^k \in \mathbb{R}^{\mathbb{M}}, i = 1, \ldots, N_k\}$ as training samples for class C_k where N_k is the number of training samples for the class C_k and $N = \sum_{k=1}^{K} N_k$ is the total number of training samples.
2) Take $\{\mu_q^k | \mu_q^k \ in \mathbb{R}^{\mathbb{M}}, q = 1, \ldots, Q\}$ as intial centroids for class C_k where $Q < N_K$ is the number of centroids for class C_k.

3)

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{19}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(19)", not "Eq. (19)" or "equation (19)", except at the beginning of a sentence: "Equation (19) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BIBTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BIBTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks …”. Instead, try “R. B. G. thanks…”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first …”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.