# MULTIVARIATE OPTIMIZATION FOR NON-DECOMPOSABLE PERFORMANCE MEASURES

Debojyoti Dey (15511264)
Nimisha Agarwal (15511267)
(Group 15)

November 15, 2016

# Overview

# Problem Statement

Finding general optimization technique to maximize performance which are non-decomposable in nature. Our work will primarily consider some concave measures of performance.

- Examples: Min, G-mean, H-mean etc.
- Expressed as $f(TPR, TNR)$.
- TPR = True Positive Rate, TNR = True Negative Rate

# Decomposable vs Non-decomposable

- Misclassification rate is decomposable. Loss function:

$$\Delta(\bar{y}, \bar{y}^*) = \sum_{i=1}^{n} \frac{1 - y_i y_i^*}{2}$$

- $F_\beta$ score is non-decomposable. Performance measure:

$$\phi(\bar{y}, \bar{y}^*) = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP}$$

- Problem with misclassification rate?
    - class imbalanced setting

# Multivariate Optimization Setting

- Hypothesis function $\overline{h}$ maps $\overline{x} \in \overline{\mathcal{X}}$ and $\overline{x} = \{x_1, x_2, \cdots, x_n\}$ to $\overline{y} \in \overline{\mathcal{Y}}$ and $\overline{y} = \{y_1, y_2, \cdots, y_n\}$ where $y_i \in \{+1, -1\}$

$$\overline{h} : \overline{\mathcal{X}} \rightarrow \overline{\mathcal{Y}}$$

- We define score for a particular input-output combination as follows:

$$f_w(\overline{x}, \overline{y}) = w^T \psi(\overline{x}, \overline{y})$$

- Hypothesis function gives $\overline{y}$ with highest score for an input $\overline{x}$.

$$\overline{h}_w(\overline{x}) = \arg\max_{\overline{y} \in \mathcal{Y}} (w^T \psi(\overline{x}, \overline{y}))$$

- We use the following form of $\psi$

$$\psi(\overline{x}, \overline{y}) = \sum_{i=1}^{n} y_i^* x_i$$

# Structural SVM

- Struct SVM by [Joachims:2005] used for multi-class classification.

$$
\begin{aligned}
&\min_{w,\xi\geq 0} \quad \frac{1}{2}\|w\|^2 + C\xi \\
&s.t \quad \forall \overline{y} \in \overline{\mathcal{Y}} \setminus \overline{y}^* : w^T[\psi(\overline{x},\overline{y}^*) - \psi(\overline{x},\overline{y})] \geq \Delta(\overline{y}^*,\overline{y}) - \xi \\
&\Rightarrow \Delta(\overline{y}^*,\overline{y}) + \Sigma(y_i - y_i^*)w^T x_i \leq \xi
\end{aligned}
$$

  where $\Delta(\overline{y}^*,\overline{y})$ is loss function.
- $\xi$ is the upper bound of loss function.
- We substitute margin violation $\xi$ in objective function by

$$
\mathcal{L}_w(\overline{x},\overline{y}^*) = \max_{\overline{y}\in\{1,-1\}^n} \{\Delta(\overline{y}^*,\overline{y}) + \sum_{i=1}^{n}(y_i - y_i^*)w^T x_i\} \qquad (1)
$$

# Performance measure in Fenchel Dual

- Performance measure G-mean is given by

$$\begin{aligned}\phi(P, N) &= \sqrt{PN} \\ &= \min_{\alpha,\beta}\{\alpha P + \beta N - \phi^*(\alpha, \beta)\}\end{aligned}$$

  as $\phi$ is a concave function. P,N stands for TPR and TNR respectively.
- We define our loss function as

$$\begin{aligned}\Delta(\bar{y}^*, \bar{y}) &= -\phi(P, N) \\ &= \max_{\alpha,\beta}\{-\alpha P - \beta N + \phi^*(\alpha, \beta)\} \\ &= \max_{\alpha,\beta}\{\alpha P + \beta N - (-\phi^*(-\alpha, -\beta))\}\end{aligned}$$

# Optimization Problem

- We can re-write equation 1 as,

$$\mathcal{L}_w(\overline{x}, \overline{y}^*)$$

$$= \max_{\overline{y} \in \{1,-1\}^n} \{\max_{\alpha,\beta}\{\alpha P + \beta N + \phi^*(-\alpha, -\beta)\} + \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)w^T x_i\}$$

$$= \max_{\alpha,\beta}\{\max_{\overline{y}}\{\alpha P + \beta N + \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)w^T x_i\} + \phi^*(-\alpha, -\beta)\}$$

- Express $P$ and $N$ as following:

$$P = \sum_{i=1}^{n} P_i(y_i, y_i^*) = \frac{1}{n_+}\sum_{i=1}^{n}\frac{(1 + y_i)(1 + y_i^*)}{4}$$

# Continued...

$$N = \sum_{i=1}^{n} N_i(y_i, y_i^*) = \frac{1}{n_-} \sum_{i=1}^{n} \frac{(1 - y_i)(1 - y_i^*)}{4}$$

- Substituting, inner maximization becomes,

$$\sum_{i=1}^{n} \max_{y_i \in \{-1,+1\}} \frac{\alpha}{n_+} \frac{(1 + y_i)(1 + y_i^*)}{4} + \frac{\beta}{n_-} \frac{(1 - y_i)(1 - y_i^*)}{4} + \frac{1}{n}(y_i - y_i^*) w^T x_i$$

  using independence among the data points.
- Now we can perform maximization for each point separately.

# Continued...

Solving the above maximization, we get the following hinge loss like function,

$$\frac{\alpha}{n_+}\max\{0, 1 - y_i^* \frac{2n_+}{\alpha n} w^T x_i\} - \frac{\alpha}{n_+}$$

for positive points and similarly for negatives.

# Shortcoming of SPADE

- Requires link functions to be L-Lipschitz.
- Works with non-Lipschitz functions when some restrictions are imposed on them.

# References


Joachims, Thorsten
A Support Vector Method for Multivariate Performance Measures.
*Proceedings of the 22Nd International Conference on Machine Learning*

Narasimhan, Harikrishna and Vaish, Rohit and Agarwal, Shivani
On the Statistical Consistency of Plug-in Classifiers for Non-decomposable Performance Measures.
*Advances in Neural Information Processing Systems 27*

Narasimhan, Harikrishna and Kar, Purushottam and Jain, Prateek
Optimizing Non-decomposable Performance Measures: A Tale of Two Classes.
*ICML*

Questions?