MULTIVARIATE OPTIMIZATION FOR NON-DECOMPOSABLE PERFORMANCE MEASURES

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

November 17, 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

Stochastic Primal Dual Method [NarasimhanK015]

The existing online method has following shortcomings:

- Requires performance measure function to be L-Lipschitz.
- Works with non-Lipschitz functions with some restriction.
- Example: G-mean

Decomposable vs Non-decomposable

- Misclassification rate is decomposable. Loss function:

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

- F_β score is non-decomposable. Performance measure:

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

- 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 $y_i \in \{+1, -1\}$

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

- 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 ψ

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

Structural SVM

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

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

where $\Delta(\overline{y}^*,\overline{y})$ is loss function.

- ξ is the upper bound of loss function.
- We substitute margin violation ξ in objective function by

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

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 ϕ is a concave function. P,N stands for TPR and TNR respectively.

- We define our loss function as

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

Optimization Problem

- We can re-write equation 1 as,

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

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

$$= \max_{\alpha,\beta}\{\max_{\overline{y}}\{-\alpha P - \beta N + \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)w^T x_i\} + \phi^*(\alpha, \beta)\} \quad (2)$$

- 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 weighted hinge loss like function with some additional constants,

$$\sum_{i=1}^{n}\left(\frac{\alpha}{n_+}\max\{0, 1 - y_i^*\frac{2n_+}{\alpha n}w^T x_i\} - \frac{\alpha}{n_+}\right)\mathbb{I}(y_i^* = 1)$$
$$+\left(\frac{\beta}{n_-}\max\{0, 1 - y_i^*\frac{2n_-}{\beta n}w^T x_i\} - \frac{\beta}{n_-}\right)\mathbb{I}(y_i^* = -1)$$

Now we can re-write equation 2 as following

$$\begin{aligned}\mathcal{L}_w(\overline{x}, \overline{y}^*) = \max_{\alpha,\beta}\{&\frac{\alpha}{n_+}\sum_{y_i^*=1}\max\{0, 1 - y_i^*\frac{2n_+}{\alpha n}w^T x_i\} \\ &+ \frac{\beta}{n_-}\sum_{y_i^*=-1}\max\{0, 1 - y_i^*\frac{2n_-}{\beta n}w^T x_i\} \\ &- (\alpha + \beta) + \phi^*(\alpha, \beta)\}\end{aligned}$$

Objective function

We can substitute the loss in struct SVM:

$$
\begin{aligned}
&\min_w \frac{||w||^2}{2} + C\mathcal{L}_w(\overline{x}, \overline{y}^*) \\
\equiv &\min_w \frac{||w||^2}{2} + C \max_{\alpha,\beta}\{ \frac{\alpha}{n_+} \sum_{y_i^*=1} \max\{0, 1 - y_i^* \frac{2n_+}{\alpha n} w^T x_i\} \\
&\qquad + \frac{\beta}{n_-} \sum_{y_i^*=-1} \max\{0, 1 - y_i^* \frac{2n_-}{\beta n} w^T x_i\} \\
&\qquad - (\alpha + \beta) + \phi^*(\alpha, \beta)\}
\end{aligned}
$$

Solution Steps

We perform the following steps in each iterative cycle:

1. Fix w
2. Gradient ascent on $\mathcal{L}_w(\overline{x}, \overline{y}^*)$ wrt α, β
3. Fix α, β
4. SVM wrt w
5. Go to step 1

We can use Liblinear solver to perform SVM.

Conjugate function of $\phi(P, N)$

For any concave ϕ

$$\phi(P, N) = \min_{\alpha,\beta}\{\alpha P + \beta N - \phi^*(\alpha, \beta)\}$$

For G-mean as ϕ

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

Solving for $g'(P, N) = 0$ where $g = \alpha P + \beta N - \sqrt{PN}$, we get $\alpha = \frac{1}{2}\sqrt{\frac{N}{P}}$ and $\beta = \frac{1}{2}\sqrt{\frac{P}{N}}$ giving

$$\phi^*(\alpha, \beta) = 0 \quad (3)$$

Dual feasible region

Trivial: $\alpha > 0, \beta > 0$ as $P, N \geq 0$

$$g = \alpha P + \beta N - \sqrt{PN}$$
$$= (\sqrt{\alpha P} - \sqrt{\beta N})^2 + \sqrt{PN}(2\sqrt{\alpha\beta} - 1)$$

By ensuring $P = \frac{\beta N}{\alpha}$ we can show the first part to be zero. For large P, N, $g \to -\infty$ if $2\sqrt{\alpha\beta} < 1$. Hence the feasible dual region is defined by,

$$dom(\alpha, \beta) = \{\alpha, \beta | \alpha\beta \geq \frac{1}{4}, \alpha > 0, \beta > 0\}$$

$\mathbb{Q}$. P, N has to be the number of true positives/negatives, not rate?

Gradient computation

- $h(\alpha) = \frac{\alpha}{n_+} \max\{0, 1 - y_i^* \frac{2n_+}{\alpha n} w^T x_i\}$. By Danskin's theorem it can be shown that,

$$h'(\alpha) = \begin{cases} 0, & \text{if } y_i^* \frac{2n_+}{\alpha n} w^T x_i \geq 1 \\ \frac{1}{n_+}, & \text{otherwise} \end{cases}$$

- Similarly for $h(\beta) = \frac{\beta}{n_-} \max\{0, 1 - y_i^* \frac{2n_-}{\alpha n} w^T x_i\}$. By Danskin's theorem it can be shown that,

$$h'(\beta) = \begin{cases} 0, & \text{if } y_i^* \frac{2n_-}{\beta n} w^T x_i \geq 1 \\ \frac{1}{n_-}, & \text{otherwise} \end{cases}$$

- We have proved $\phi^*(a, b) = 0$ for G-mean, hence its derivative gives 0.

Problem with Gradient Ascent

- Unbounded increase in dual variables in $\mathbb{R}^2_+$

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

Thank You!