Generalization Error Bounds for Learning to Rank

Ambuj Tewari and Sougata Chaudhuri

Department of Statistics, and
Department of EECS,
University of Michigan, Ann Arbor

June 28, 2015

Learning to Rank at Query Level

Learning to Rank at Query Level

- Input space- Lists of m documents pertaining to queries.
  - Formally: $\mathcal{X} \in \mathbb{R}^{m \times d}$

Learning to Rank at Query Level

- Input space- Lists of m documents pertaining to queries.
  - Formally: $\mathcal{X} \in \mathbb{R}^{m \times d}$
- Supervision space- Relevance vectors of length m.
  - Formally: $\mathcal{Y} \in \{0, 1, \ldots, K\}^m$.

Learning to Rank at Query Level

- Input space- Lists of m documents pertaining to queries.
  - Formally: $\mathcal{X} \in \mathbb{R}^{m \times d}$
- Supervision space- Relevance vectors of length m.
  - Formally: $\mathcal{Y} \in \{0, 1, \ldots, K\}^m$.
- Rank documents by sorting scores corresponding to a scoring function.

Learning to Rank at Query Level

- Input space- Lists of m documents pertaining to queries.
  - Formally: $\mathcal{X} \in \mathbb{R}^{m \times d}$
- Supervision space- Relevance vectors of length m.
  - Formally: $\mathcal{Y} \in \{0, 1, \ldots, K\}^m$.
- Rank documents by sorting scores corresponding to a scoring function.
- For $X \in \mathcal{X}$, *linear* scoring function: $f_w(X) = Xw \in \mathbb{R}^m$.

Ranking Surrogates

- Scoring function learnt from training data.
  - Training data: $(X_i, R_i) \overset{i.i.d}{\sim} \mathbb{D}(\mathcal{X} \times \mathcal{Y})$

Ranking Surrogates

- Scoring function learnt from training data.
  - Training data: $(X_i, R_i) \overset{i.i.d}{\sim} \mathbb{D}(\mathcal{X} \times \mathcal{Y})$
- Performance of function judged by target measures like NDCG, AP.

Ranking Surrogates

- Scoring function learnt from training data.
  - Training data: $(X_i, R_i) \overset{i.i.d}{\sim} \mathbb{D}(\mathcal{X} \times \mathcal{Y})$
- Performance of function judged by target measures like NDCG, AP.
- Computationally difficult to optimize the measures during training time.

Ranking Surrogates

- Scoring function learnt from training data.
  - Training data: $(X_i, R_i) \overset{i.i.d}{\sim} \mathbb{D}(\mathcal{X} \times \mathcal{Y})$
- Performance of function judged by target measures like NDCG, AP.
- Computationally difficult to optimize the measures during training time.
- Hence, development of a number of ranking surrogates.

Generalization and Calibration

- ERM algorithms learn a function by minimizing surrogate loss on training data.

Generalization and Calibration

- ERM algorithms learn a function by minimizing surrogate loss on training data.
- **Generalization Error**: What is the expected surrogate loss for the learnt function?

Generalization and Calibration

- ERM algorithms learn a function by minimizing surrogate loss on training data.
- **Generalization Error**: What is the expected surrogate loss for the learnt function?
- **Calibration**: How does expected surrogate loss relate to expected *target measures* based losses?

Generalization and Calibration

- ERM algorithms learn a function by minimizing surrogate loss on training data.
- **Generalization Error**: What is the expected surrogate loss for the learnt function?
- **Calibration**: How does expected surrogate loss relate to expected *target measures* based losses?
- We address question on **generalization error**.

Generalization Error Equation

- Let ϕ be a ranking surrogate.
  - $\phi(s^w, R) \mapsto \mathbb{R}$, for $\underbrace{s^w}_{\text{score}} = Xw \in \mathbb{R}^m$, $\underbrace{R}_{\text{relevance}} \in \{0, \ldots, K\}^m$.

Generalization Error Equation

- Let ϕ be a ranking surrogate.
  - $\phi(s^w, R) \mapsto \mathbb{R}$, for $\underbrace{s^w}_{\text{score}} = Xw \in \mathbb{R}^m$, $\underbrace{R}_{\text{relevance}} \in \{0, \ldots, K\}^m$.
- Uniform generalization error (over parameter class $\mathcal{F}$):

$$\underbrace{\mathbb{E}[\phi(s^w, R)]}_{\text{expected loss}} \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \phi(s_i^w, R_i)}_{\text{empirical loss}} + \mathit{Complexity}, \quad \forall\ w \in \mathcal{F}.$$

Generalization Error Equation

- Let ϕ be a ranking surrogate.
  - $\phi(s^w, R) \mapsto \mathbb{R}$, for $\underbrace{s^w}_{\text{score}} = Xw \in \mathbb{R}^m$, $\underbrace{R}_{\text{relevance}} \in \{0, \ldots, K\}^m$.
- Uniform generalization error (over parameter class $\mathcal{F}$):

$$\underbrace{\mathbb{E}[\phi(s^w, R)]}_{\text{expected loss}} \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \phi(s_i^w, R_i)}_{\text{empirical loss}} + Complexity, \quad \forall\ w \in \mathcal{F}.$$

- Complexity term may depend on properties of ϕ, sample size n, length of document list m etc.

Lipschitz Surrogates

- Since $\phi(s, R)$ is defined on vector valued predictions ($s \in \mathbb{R}^m$), Lipschitz property dependent on norm.

Lipschitz Surrogates

- Since $\phi(s, R)$ is defined on vector valued predictions ($s \in \mathbb{R}^m$), Lipschitz property dependent on norm.

- $\phi(s, R)$ is L_2 Lipschitz w.r.t s in ℓ_2 norm if $|\phi(s_1, R) - \phi(s_2, R)| \leq L_2 \|s_1 - s_2\|_2$.

Lipschitz Surrogates

- Since $\phi(s, R)$ is defined on vector valued predictions ($s \in \mathbb{R}^m$), Lipschitz property dependent on norm.
- $\phi(s, R)$ is L_2 Lipschitz w.r.t s in ℓ_2 norm if $|\phi(s_1, R) - \phi(s_2, R)| \leq L_2 \|s_1 - s_2\|_2$.
- $\phi(s, R)$ is L_1 Lipschitz w.r.t s in ℓ_∞ norm if $|\phi(s_1, R) - \phi(s_2, R)| \leq L_2 1 \|s_1 - s_2\|_\infty$.

Lipschitz Surrogates

- Since $\phi(s, R)$ is defined on vector valued predictions ($s \in \mathbb{R}^m$), Lipschitz property dependent on norm.
- $\phi(s, R)$ is L_2 Lipschitz w.r.t s in ℓ_2 norm if $|\phi(s_1, R) - \phi(s_2, R)| \leq L_2 \|s_1 - s_2\|_2$.
- $\phi(s, R)$ is L_1 Lipschitz w.r.t s in ℓ_∞ norm if $|\phi(s_1, R) - \phi(s_2, R)| \leq L_2 1 \|s_1 - s_2\|_\infty$.
- $L_1 \leq \sqrt{m} L_2$.

Existing Result

Existing Result

- Let $\|w\|_2 \leq W_2$, R_X be bound on ℓ_2 norm of feature vectors.

Existing Result

- Let $\|w\|_2 \leq W_2$, R_X be bound on ℓ_2 norm of feature vectors.
- Best known complexity for Lipschitz surrogates in ℓ_2 norm: $O(L_2 W_2 R_X \sqrt{\frac{m}{n}})$.

Existing Result

- Let $\|w\|_2 \leq W_2$, R_X be bound on ℓ_2 norm of feature vectors.
- Best known complexity for Lipschitz surrogates in ℓ_2 norm: $O(L_2 W_2 R_X \sqrt{\frac{m}{n}})$.
- Proof technique was intrinsic to ℓ_2 Lipschitz surrogates and necessitated m dependence.

Necessary Dependence on m?

- Complexity term depends on richness of class of scoring functions.

Necessary Dependence on m?

- Complexity term depends on richness of class of scoring functions.
- Linear scoring functions parameterized by d dimensional vector ($w \in \mathbb{R}^d$), *independent of* m.

Necessary Dependence on m?

- Complexity term depends on richness of class of scoring functions.
- Linear scoring functions parameterized by d dimensional vector ($w \in \mathbb{R}^d$), *independent of* m.
- Should complexity term be independent of m?

Necessary Dependence on m?

- Complexity term depends on richness of class of scoring functions.
- Linear scoring functions parameterized by d dimensional vector ($w \in \mathbb{R}^d$), *independent of* m.
- Should complexity term be independent of m?
- What role does Lipschitz norm play in complexity?

Our Contributions

Examples- ListNet and SmoothDCG

Examples- ListNet and SmoothDCG

- Listnet (convex) and SmoothDCG@1 (non-convex) are popular ranking surrogates.

Examples- ListNet and SmoothDCG

- Listnet (convex) and SmoothDCG@1 (non-convex) are popular ranking surrogates.
- Both are ℓ_∞ Lipschitz with constants independent of m

Examples- ListNet and SmoothDCG

- Listnet (convex) and SmoothDCG@1 (non-convex) are popular ranking surrogates.
- Both are ℓ_∞ Lipschitz with constants independent of m
- Previous generalization error bounds for both the surrogates had m dependent complexity.

Results for Convex and Lipschitz Surrogates

ϕ is ℓ_∞ Lipschitz (constant L_1), functional parameter $\|w\|_2 \leq W_2$, R_X bound on ℓ_2 norm of feature vectors.

Methods	Non-convexity	Complexity	Constants in $O(\cdot)$
OGD	No	$O(L_1 W_2 R_X \sqrt{\frac{1}{n}})$	smallest
RERM	No	$O(L_1 W_2 R_X \sqrt{\frac{1}{n}})$	small
ERM	Yes	$O(L_1 W_2 R_X \sqrt{\frac{1}{n}})$	several log factors

Analysis of Methods

Analysis of Methods

- OGD: Specific to convex surrogates and Online Gradient Descent algorithm.

Analysis of Methods

- OGD: Specific to convex surrogates and Online Gradient Descent algorithm.
- RERM: Specific to convex surrogates, requires ℓ_2 regularization function.

Analysis of Methods

- OGD: Specific to convex surrogates and Online Gradient Descent algorithm.
- RERM: Specific to convex surrogates, requires ℓ_2 regularization function.
- ERM: Applies to all Lipschitz surrogates, for all Empirical Risk Minimization algorithms.

Result for High Dimensional Feature

- Learning to rank problems can involve high dimensional features.

Result for High Dimensional Feature

- Learning to rank problems can involve high dimensional features.
- Appropriately, let $\{w \in \mathbb{R}^d : \|w\|_1 \leq W_1\}$, $\bar{R}_X$ be bound on ℓ_∞ norm of feature vectors.

Result for High Dimensional Feature

- Learning to rank problems can involve high dimensional features.
- Appropriately, let $\{w \in \mathbb{R}^d : \|w\|_1 \leq W_1\}$, $\bar{R}_X$ be bound on ℓ_∞ norm of feature vectors.
- Generalization error complexity: $O(L_1 W_1 \bar{R}_X \sqrt{\frac{\log(d)}{n}})$.

Result for High Dimensional Feature

- Learning to rank problems can involve high dimensional features.
- Appropriately, let $\{w \in \mathbb{R}^d : \|w\|_1 \leq W_1\}$, $\bar{R}_X$ be bound on ℓ_∞ norm of feature vectors.
- Generalization error complexity: $O(L_1 W_1 \bar{R}_X \sqrt{\frac{\log(d)}{n}})$.
- Complexity *nearly independent* of d.

Result for Smooth Surrogates

- Let ϕ be a smooth surrogate w.r.t ℓ_∞ norm with constant H_ϕ(definition in paper).

Result for Smooth Surrogates

- Let ϕ be a smooth surrogate w.r.t ℓ_∞ norm with constant H_ϕ(definition in paper).
- Let $L_\phi(w^*) = \min_w \mathbb{E}[\phi(s^w, R)]$ and C be constant depending on H_ϕ.

Result for Smooth Surrogates

- Let ϕ be a smooth surrogate w.r.t ℓ_∞ norm with constant H_ϕ(definition in paper).
- Let $L_\phi(w^*) = \min_w \mathbb{E}[\phi(s^w, R)]$ and C be constant depending on H_ϕ.
- Generalization error complexity: $O(\sqrt{\frac{L_\phi(w^*)C}{n}} + \frac{C}{n})$.

Result for Smooth Surrogates

- Let ϕ be a smooth surrogate w.r.t ℓ_∞ norm with constant H_ϕ(definition in paper).
- Let $L_\phi(w^*) = \min_w \mathbb{E}[\phi(s^w, R)]$ and C be constant depending on H_ϕ.
- Generalization error complexity: $O(\sqrt{\frac{L_\phi(w^*)C}{n}} + \frac{C}{n})$.
- Rate interpolates between $O(\frac{1}{n})$ $(L_\phi(w^*) = 0)$ and $O(\sqrt{\frac{1}{n}})$ $(L_\phi(w^*) > 0)$.