Toxic comment classification challenge

Руднев Виктор, Солоткий Михаил

МГУ им. Ломоносова, факультет ВМК, кафедра ММП

22 марта 2018 г.

Датасет

- Комментарии из правок к Википедии
- В train set $\approx$ 150000 комментариев.
- Теги очень не сбалансированы (clean — 135k, toxic — 15k, threat — 1k)
- Мера качества: средний roc-auc по 6 target'ам. (очень плохая метрика)

---

https://www.kaggle.com/lct14558/imbalanced-data-why-you-should-not-use-roc-curve

Baseline

- MultinomialNB без tf-idf
  - private: 0.9084, public: 0.9091
- MultinomialNB
  - private: 0.9636, public: 0.9601
- LogisticRegression с char n-grams
  - private: 0.9804, public: 0.9790

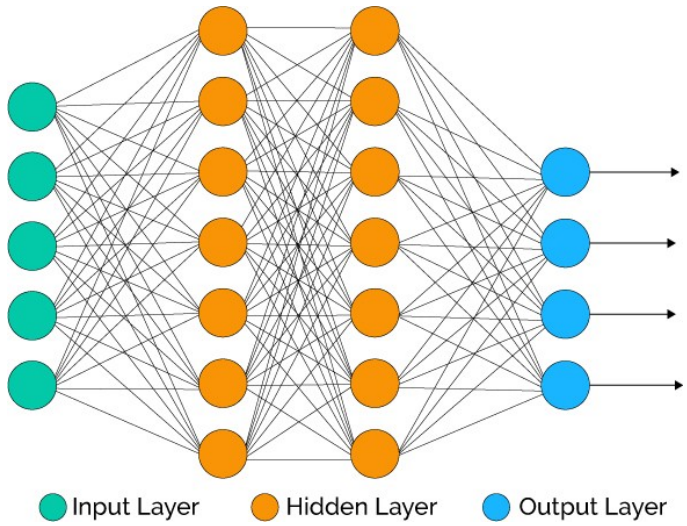
Первый бленд

- rank averaged submission
  private: 0.9850, public: 0.9855

---

https://www.kaggle.com/fizzbuzz/rank-averaging-on-preprocessed-data

Сложный выбор



Второй бленд

- Another Cleaned Data Blend With Low Correlation
  private: 0.9853, public: 0.9860

---

https://www.kaggle.com/antmarakis/another-cleaned-data-blend-with-low-correlation

Третий бленд

- Blend It All
  private: 0.9863, public: 0.9867

---

https://www.kaggle.com/rednivrug/blend-it-all

Наше улучшение

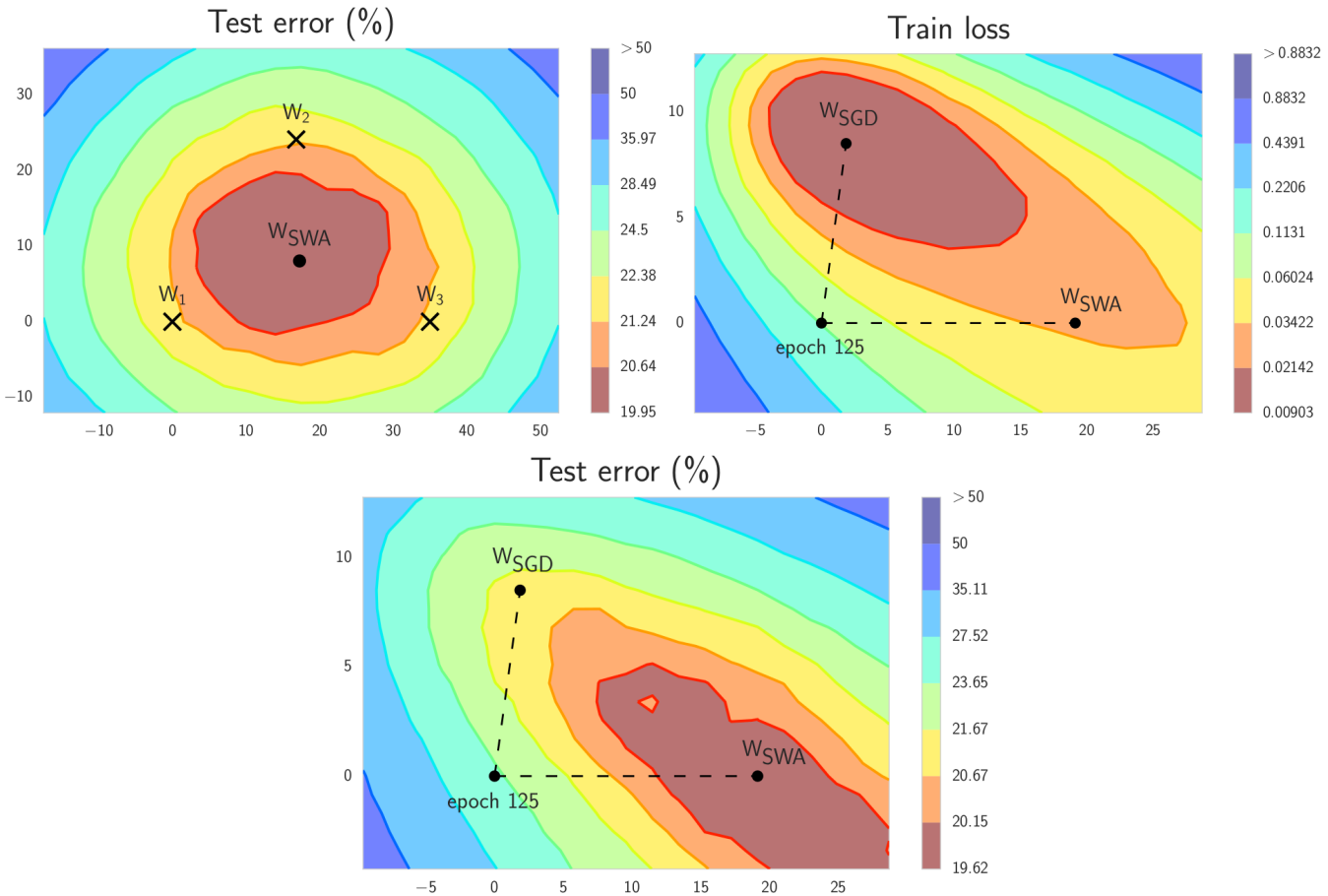
- $\text{total} = \text{blend\_it\_all} \cdot 12 - \text{logreg}$
  С примерно 950 места на 456
  private: 0.9863, public: 0.9867

Сетки

- Glove -> Bi-GRU -> Conv1d -> Dense
  - private: 0.9826, public: 0.9840
  - private: 0.9832, public: 0.9847 (прокачанный)
- Glove -> Bi-GRU -> MaxPool -> Dense
  - private: 0.9811, public: 0.9823
- Fasttext -> 2xBi-GRU -> MaxPool -> 2xDense
  - private: 0.9818, public: 0.9833

(и тонна Dropout-ов повсюду)

Stochastic Weight Averaging



https://arxiv.org/abs/1803.05407

Stochastic Weight Averaging

Require:
weights $\hat{w}$, number of iterations n
Ensure: w_{SWA}
$w \leftarrow \hat{w}$ {Initialize weights with $\hat{w}$}
$w_{\text{SWA}} \leftarrow w$
for $i \leftarrow 1, 2, \ldots, n$ **do**
$\alpha \leftarrow \alpha(i)$ {Calculate LR for the iteration}
$w \leftarrow w - \alpha \nabla \mathcal{L}_i(w)$ {SGD update}
$n_{\text{models}} \leftarrow i$ {Number of models}
$w_{\text{SWA}} \leftarrow \frac{w_{\text{SWA}} \cdot n_{\text{models}} + w}{n_{\text{models}} + 1}$ {Update average}
end for

Stochastic Weight Averaging

- Adam
  - private: 0.9826, public: 0.9840
- Бленд 4 запусков SGD
  - private: 0.9830, public: 0.9842
- SWA
  - private: 0.9832, public: 0.9847

Blending is the new sexy!

---

https://www.kaggle.com/jagangupta/lessons-from-toxic-blending-is-the-new-sexy

Ссылки I

Izmailov, Pavel and Podoprikhin, Dmitrii and Garipov, Timur and Vetrov, Dmitry and Wilson, Andrew Gordon

Averaging Weights Leads to Wider Optima and Better Generalization

arXiv preprint, 2018,

https://arxiv.org/abs/1803.05407

Zafar

Rank averaging on preprocessed data

https://www.kaggle.com/fizzbuzz/rank-averaging-on-preprocessed-data

Ссылки II

Anthony Marakis

Another cleaned data blend with low correlation

https://www.kaggle.com/antmarakis/another-cleaned-data-blend-with-low-correlation

rednivrug

Blend it all

https://www.kaggle.com/rednivrug/blend-it-all

Ashish Gupta

Bidirectional GRU with Convolution

https://www.kaggle.com/eashish/bidirectional-gru-with-convolution

Ссылки III

Prashant Kikani

Pooled GRU + GloVe (with preprocessing)

https://www.kaggle.com/prashantkikani/pooled-gru-glove-with-preprocessing

zhbain

Pooled GRU + FastText

https://www.kaggle.com/zhbain/pooled-gru-fasttext-6c07c9

Ссылки IV

Jagan

Lessons from Toxic : Blending is the new sexy

https://www.kaggle.com/jagangupta/lessons-from-toxic-blending-is-the-new-sexy

LCT14558

Imbalanced data why you should NOT use ROC curve

https://www.kaggle.com/lct14558/imbalanced-data-why-you-should-not-use-roc-curve

Ссылки V

Git репозиторий

`https://github.com/MichaelSolotky/ToxicCommentsKaggle/tree/master`