PAC-Bayes оценки и их приложения в машинном обучении

И. О. Толстихин, ВЦ РАН
iliya.tolstikhin@gmail.com

октябрь 2012

Основная тема доклада — теория обобщающей способности

- Какие гарантии о качестве настраиваемого по обучающей выборке алгоритма мы можем получать?
- Какие ограничения на свойства рассматриваемой задачи при этом нужны?
- Применимы ли такие результаты на практике?
- Если да — то как их применять?

Основная тема доклада — теория обобщающей способности

- Какие гарантии о качестве настраиваемого по обучающей выборке алгоритма мы можем получать?
- Какие ограничения на свойства рассматриваемой задачи при этом нужны?
- Применимы ли такие результаты на практике?
- Если да — то как их применять?

Основная тема доклада — теория обобщающей способности

- Какие гарантии о качестве настраиваемого по обучающей выборке алгоритма мы можем получать?
- Какие ограничения на свойства рассматриваемой задачи при этом нужны?
- Применимы ли такие результаты на практике?
- Если да — то как их применять?

Основная тема доклада — теория обобщающей способности

- Какие гарантии о качестве настраиваемого по обучающей выборке алгоритма мы можем получать?
- Какие ограничения на свойства рассматриваемой задачи при этом нужны?
- Применимы ли такие результаты на практике?
- Если да — то как их применять?

PAC-Bayes подход:

«Одни из самых точных оценок обобщающей способности».

«Одни из самых простых доказательств оценок».

Содержание

1. **Введение**
   - Теория статистического обучения
   - Байесовский подход
   - PAC-Bayes подход к переобучению
2. **PAC-Bayes оценки**
   - Определения и постановка задачи
   - Основные теоремы PAC-Bayes
3. **Улучшения и приложения PAC-Bayes**
   - Приложения: линейные классификаторы
   - Выбор априорного распределения

Определения: Statistical Learning Theory (SLT)

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*.

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g \colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r \colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$. Пусть $r \in [0, 1]$.
(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$Pg \stackrel{\text{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$P_\ell g \stackrel{\text{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

Определения: Statistical Learning Theory (SLT)

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*.

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g \colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r \colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^{+}$. Пусть $r \in [0, 1]$.
(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$Pg \stackrel{\mathrm{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$P_\ell g \stackrel{\mathrm{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

Определения: Statistical Learning Theory (SLT)

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*.

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g \colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r \colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^{+}$. Пусть $r \in [0, 1]$.
(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$Pg \stackrel{\text{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$P_\ell g \stackrel{\text{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

Определения: Statistical Learning Theory (SLT)

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*.

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g\colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^{+}$. Пусть $r \in [0, 1]$.

(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$Pg \stackrel{\text{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$P_\ell g \stackrel{\text{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

Определения: Statistical Learning Theory (SLT)

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*.

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g\colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$. Пусть $r \in [0, 1]$.
(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$Pg \stackrel{\text{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$P_\ell g \stackrel{\text{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

Определения: Statistical Learning Theory (SLT)

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*.

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g\colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^{+}$. Пусть $r \in [0, 1]$.
(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$Pg \stackrel{\text{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\bigl(g(X), Y\bigr).$$

Эмпирический риск классификатора:

$$P_\ell g \stackrel{\text{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\bigl(g(X_i), Y_i\bigr).$$

Границы применимости SLT

Мы ввели следующие ограничения:

- Распределение P фиксировано и неизвестно;
- Обучающая выборка — независимые, одинаково распределенные с. в. из P.
- Все выборки, которые мы можем наблюдать в будущем, также простые из P.

Этап обучения

Мы хотим минимизировать риск:

$$Pg \stackrel{\text{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big) \to \min_{g\in\mathcal{G}}.$$

Но распределение P неизвестно, поэтому мы будем минимизировать эмпирический риск:

$$P_\ell g \stackrel{\text{def}}{=} \frac{1}{\ell}\sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big) \to \min_{g\in\mathcal{G}}. \qquad \text{(МЭР)}$$

Решение задачи (МЭР) — g^ℓ.

Основные задачи SLT

- Изучение поведения случайной величины Pg^{ℓ}.
  - generalization bounds:
    относительно ее эмпирического риска $P_{\ell}g^{\ell}$;
  - excess risk bounds:
    относительно риска лучшего в классе $\mathcal{G}$ классификатора $g^* = \arg\min_{g\in\mathcal{G}} Pg$;
  - ...
- Выбор *оптимального* множества $\mathcal{G}$ (model selection).
  - Structural Risk Minimization;
  - Oracle inequalities;
  - Complexity penalization;
  - ...

Основные задачи SLT

- Изучение поведения случайной величины Pg^ℓ.
  - generalization bounds:
    относительно ее эмпирического риска $P_\ell g^\ell$;
  - excess risk bounds:
    относительно риска лучшего в классе $\mathcal{G}$ классификатора $g^* = \arg\min_{g \in \mathcal{G}} Pg$;
  - ...
- Выбор *оптимального* множества $\mathcal{G}$ (model selection).
  - Structural Risk Minimization;
  - Oracle inequalities;
  - Complexity penalization;
  - ...

Основные задачи SLT

- Изучение поведения случайной величины Pg^{ℓ}.
    - generalization bounds:
      относительно ее эмпирического риска $P_{\ell} g^{\ell}$;
    - excess risk bounds:
      относительно риска лучшего в классе $\mathcal{G}$ классификатора $g^* = \arg\min_{g \in \mathcal{G}} Pg$;
    - ...
- Выбор *оптимального* множества $\mathcal{G}$ (model selection).
    - Structural Risk Minimization;
    - Oracle inequalities;
    - Complexity penalization;
    - ...

Основные задачи SLT

- Изучение поведения случайной величины Pg^{ℓ}.
  - generalization bounds:
    относительно ее эмпирического риска $P_{\ell}g^{\ell}$;
  - excess risk bounds:
    относительно риска лучшего в классе $\mathcal{G}$ классификатора $g^* = \arg\min_{g \in \mathcal{G}} Pg$;
  - ...
- Выбор *оптимального* множества $\mathcal{G}$ (model selection).
  - Structural Risk Minimization;
  - Oracle inequalities;
  - Complexity penalization;
  - ...

Классические результаты: VC-теория

Для всех $g \in \mathcal{G}$ с вероятностью не меньше $1 - \delta$ относительно случайных реализаций обучающей выборки справедливо:

$$Pg \leqslant P_\ell g + C(\mathcal{G}, \ell, \delta),$$

где C — характеристика сложности класса $\mathcal{G}$ (complexity):

- конечный класс $\mathcal{G}$:

$$C(\mathcal{G}, \ell, \delta) = \sqrt{\frac{\ln|\mathcal{G}| + \ln\frac{1}{\delta}}{2\ell}}, \text{ где } |\mathcal{G}| \text{ — можность.}$$

- бесконечный (несчетный) класс $\mathcal{G}$:

$$C(\mathcal{G}, \ell, \delta) = \sqrt{2\frac{\ln S_{\mathcal{G}}(2\ell) + \ln\frac{2}{\delta}}{\ell}}, \text{ где } S_{\mathcal{G}}(\ell) \text{ — функция роста.}$$

Классические результаты: VC-теория

Для всех $g \in \mathcal{G}$ с вероятностью не меньше $1 - \delta$ относительно случайных реализаций обучающей выборки справедливо:

$$Pg \leqslant P_\ell g + C(\mathcal{G}, \ell, \delta),$$

где C — характеристика сложности класса $\mathcal{G}$ (complexity):

- конечный класс $\mathcal{G}$:

$$C(\mathcal{G}, \ell, \delta) = \sqrt{\frac{\ln|\mathcal{G}| + \ln\frac{1}{\delta}}{2\ell}},\ \text{где } |\mathcal{G}| \text{ — можность.}$$

- бесконечный (несчетный) класс $\mathcal{G}$:

$$C(\mathcal{G}, \ell, \delta) = \sqrt{2\frac{\ln S_{\mathcal{G}}(2\ell) + \ln\frac{2}{\delta}}{\ell}},\ \text{где } S_{\mathcal{G}}(\ell) \text{ — функция роста.}$$

Классические результаты: VC-теория

Для всех $g \in \mathcal{G}$ с вероятностью не меньше $1-\delta$ относительно случайных реализаций обучающей выборки справедливо:

$$Pg \leqslant P_\ell g + C(\mathcal{G}, \ell, \delta),$$

где C — характеристика сложности класса $\mathcal{G}$ (complexity):

- конечный класс $\mathcal{G}$:

$$C(\mathcal{G}, \ell, \delta) = \sqrt{\frac{\ln|\mathcal{G}| + \ln\frac{1}{\delta}}{2\ell}},\ \text{где } |\mathcal{G}| \text{ — можность.}$$

- бесконечный (несчетный) класс $\mathcal{G}$:

$$C(\mathcal{G}, \ell, \delta) = \sqrt{2\frac{\ln S_\mathcal{G}(2\ell) + \ln\frac{2}{\delta}}{\ell}},\ \text{где } S_\mathcal{G}(\ell) \text{ — функция роста.}$$

Что гарантируется?

Для большинства обучающих выборок, вытянутых независимо из любого фиксированного распределения P, нам удалось ограничить сверху риск классификатора g^ℓ.

Что гарантируется?

Для большинства обучающих выборок, вытянутых независимо из любого фиксированного распределения P, нам удалось ограничить сверху риск классификатора g^ℓ.

Проблемы

Оценки **сильно** завышены (иногда $C(\mathcal{G}, \ell, \delta) \approx 10^8$), потому что:

Проблемы

Оценки **сильно** завышены (иногда $C(\mathcal{G}, \ell, \delta) \approx 10^8$), потому что:

- не зависят от алгоритма g^ℓ;

Проблемы

Оценки **сильно** завышены (иногда $C(\mathcal{G}, \ell, \delta) \approx 10^8$), потому что:

- не зависят от алгоритма g^ℓ;
- не зависят от обучающей выборки;
- ...

Проблемы

Оценки **сильно** завышены (иногда $C(\mathcal{G}, \ell, \delta) \approx 10^8$), потому что:

- не зависят от алгоритма g^ℓ;
- не зависят от обучающей выборки;
- ...

Значит, не применимы на практике.

Проблемы

Оценки **сильно** завышены (иногда $C(\mathcal{G}, \ell, \delta) \approx 10^8$), потому что:

- не зависят от алгоритма g^ℓ;
- не зависят от обучающей выборки;
- ...

Значит, не применимы на практике.

Уточнения оценок

Основные источники улучшений:

- теория эмпирических процессов;
- концентрационные неравенства.

Уточнения оценок

Основные источники улучшений:

- теория эмпирических процессов;
- концентрационные неравенства.

P. Bartlett, O. Bousquet, S. Mendelson. (2005)
Local Rademacher Complexities.

P. Bartlett, S. Mendelson. (2006)
Empirical Risk Minimization.

O. Bousquet, V. Koltchinskii, D. Panchenko. (2002)
Some local measures of complexity of convex hulls and generalization bounds.

V. Koltchinskii. (2006)
Local Rademacher Complexities and Oracle Inequalities in Risk Minimization.

V. Koltchinskii, D. Panchenko. (2000)
Rademacher processes and bounding the risk of function learning.

P. Massart. (2000)
Some applications of concentration inequalities to statistics.

Общая идея байесовского подхода (MAP)

- Экспертно определяем априорное распределение на параметрах модели:

$$P(\boldsymbol{\theta}).$$

- Вычисляем апостериорное распределение:

$$P(\boldsymbol{\theta}|X^\ell) \sim P(X^\ell|\boldsymbol{\theta})P(\boldsymbol{\theta}).$$

- Максимизируем апостериорное распределение:

$$P(\boldsymbol{\theta}|X^\ell) \to \max_{\boldsymbol{\theta}};$$

$$\ln P(X^\ell|\boldsymbol{\theta}) + \ln P(\boldsymbol{\theta}) \to \max_{\boldsymbol{\theta}}.$$

Цель PAC-Bayes

Взять лучшее от обеих подходов:

Цель PAC-Bayes

Взять лучшее от обеих подходов:

от байесовского

- Возможность введения в модель информативного априора;

Цель PAC-Bayes

Взять лучшее от обеих подходов:

от байесовского

- Возможность введения в модель информативного априора;

от статистического

- Количественные гарантии о работе классификатора на контрольных данных;
- При этом никаких требований о правильности априора.

PAC-Bayes: Определения

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*. ($\mathbb{Y} = \{-1, +1\}$).

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g\colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$.
(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$R(g) \stackrel{\text{def}}{=} \mathrm{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$R_\ell(g) \stackrel{\text{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

PAC-Bayes: Определения

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*. ($\mathbb{Y} = \{-1, +1\}$).

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g\colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$.
(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$R(g) \stackrel{\text{def}}{=} \mathrm{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$R_\ell(g) \stackrel{\text{def}}{=} \frac{1}{\ell}\sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

PAC-Bayes: Определения

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*. ($\mathbb{Y} = \{-1, +1\}$).

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g\colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^{+}$.

(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$R(g) \stackrel{\mathrm{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$R_\ell(g) \stackrel{\mathrm{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

PAC-Bayes: Определения

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*. ($\mathbb{Y} = \{-1, +1\}$).

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g\colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$.

(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$R(g) \stackrel{\text{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$R_\ell(g) \stackrel{\text{def}}{=} \frac{1}{\ell}\sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

PAC-Bayes: Определения

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*. ($\mathbb{Y} = \{-1, +1\}$).

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g\colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$.
(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$R(g) \stackrel{\text{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$R_\ell(g) \stackrel{\text{def}}{=} \frac{1}{\ell}\sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

PAC-Bayes: Определения

$\mathbb{X}$ и $\mathbb{Y}$ — пространства *объектов* и *ответов*. ($\mathbb{Y} = \{-1, +1\}$).

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая (i.i.d) из P.

$\mathcal{G}$ — множество *классификаторов* $g \colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r \colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$.
(положим $r(y_1, y_2) = [y_1 \neq y_2]$ — классификация)

Риск классификатора:

$$R(g) \stackrel{\text{def}}{=} \mathsf{E}_{(X,Y)\sim P}\, r\big(g(X), Y\big).$$

Эмпирический риск классификатора:

$$R_\ell(g) \stackrel{\text{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\big(g(X_i), Y_i\big).$$

PAC-Bayes: случайный классификатор (Гиббса)

- Мы будем строить композицию классификаторов, выбирая *апостериорное* распределение ρ на множестве $\mathcal{G}$:

$$B_\rho(X) = \operatorname{sgn}\{\mathsf{E}_{g\sim\rho} g(X)\}.$$

- **Задача** — найти **байесовский классификатор:** композицию с минимальным риском:

$$R(B_\rho) \to \min_\rho .$$

- Вместо этого мы будем использовать случайный **классификатор Гиббса** G_ρ:
для классификации X вытянем $g \sim \rho$ и вернем $g(X)$.

PAC-Bayes: случайный классификатор (Гиббса)

- Мы будем строить композицию классификаторов, выбирая *апостериорное* распределение ρ на множестве $\mathcal{G}$:

$$B_\rho(X) = \mathrm{sgn}\big\{\mathsf{E}_{g\sim\rho} g(X)\big\}.$$

- **Задача** — найти **байесовский классификатор:** композицию с минимальным риском:

$$R(B_\rho) \to \min_\rho .$$

- Вместо этого мы будем использовать случайный **классификатор Гиббса** G_ρ: для классификации X вытянем $g \sim \rho$ и вернем $g(X)$.

PAC-Bayes: случайный классификатор (Гиббса)

- Мы будем строить композицию классификаторов, выбирая *апостериорное* распределение ρ на множестве $\mathcal{G}$:
$$B_\rho(X) = \operatorname{sgn}\bigl\{\mathsf{E}_{g\sim\rho} g(X)\bigr\}.$$
- **Задача** — найти **байесовский классификатор:** композицию с минимальным риском:
$$R(B_\rho) \to \min_\rho .$$
- Вместо этого мы будем использовать случайный **классификатор Гиббса** G_ρ:
для классификации X вытянем $g \sim \rho$ и вернем $g(X)$.

PAC-Bayes: случайный классификатор (Гиббса)

Риск классификатора Гиббса:

$$R(G_\rho) = \mathsf{E}_{g\sim\rho} R(g) = \mathsf{E}_{g\sim\rho}\mathsf{E}_{(X,Y)\sim P}[Y \neq g(X)].$$

Эмпирический риск классификатора Гиббса:

$$R_\ell(G_\rho) = \mathsf{E}_{g\sim\rho} R_\ell(g) = \mathsf{E}_{g\sim\rho}\frac{1}{\ell}\sum_{i=1}^{\ell}[Y_i \neq g(X_i)].$$

Связь рисков классификаторов B_ρ и G_ρ:

$$R(B_\rho) \leqslant 2R(G_\rho).$$

PAC-Bayes: случайный классификатор (Гиббса)

Риск классификатора Гиббса:

$$R(G_\rho) = \mathsf{E}_{g\sim\rho} R(g) = \mathsf{E}_{g\sim\rho} \mathsf{E}_{(X,Y)\sim P}[Y \neq g(X)].$$

Эмпирический риск классификатора Гиббса:

$$R_\ell(G_\rho) = \mathsf{E}_{g\sim\rho} R_\ell(g) = \mathsf{E}_{g\sim\rho} \frac{1}{\ell} \sum_{i=1}^{\ell} [Y_i \neq g(X_i)].$$

Связь рисков классификаторов B_ρ и G_ρ:

$$R(B_\rho) \leqslant 2R(G_\rho).$$

Первая теорема PAC-Bayes

Теорема (McAllester, 1998, 1999)

Пусть $\ell(y, y') \in [0, 1]$. *Зафиксируем* **любое** *априорное распределение* π *на множестве* $\mathcal{G}$. *Тогда для всех* $\delta \in (0, 1)$ *с вероятностью не меньше* $1 - \delta$ *(относительно случайного выпадения обучающей выборки) для* **всех** *апостериорных распределений* ρ **одновременно** *выплнено:*

$$R(G_\rho) \leqslant R_\ell(G_\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2\ell}},$$

где $\mathrm{KL}(\rho\|\pi) = \int \rho(x) \ln \frac{\rho(x)}{\pi(x)} dx$ *— KL-дивергенция.*

Первая теорема PAC-Bayes: доказательство

Вариационное определение KL-дивергенции:

$$\mathrm{KL}(\rho\|\pi) = \sup_f (\mathsf{E}_\rho f - \ln \mathsf{E}_\pi e^f).$$

Для всех $f : \mathcal{G} \to \mathbb{R}$ и всех пар π и ρ:

$$\mathsf{E}_\rho f \leqslant \mathrm{KL}(\rho\|\pi) + \ln \mathsf{E}_\pi e^f. \tag{v}$$

Возьмем $f = \lambda\bigl(R(g) - R_\ell(g)\bigr)$. Ограничим $\mathsf{E}_\pi e^f$:

$$\mathsf{E}_\pi e^f \leqslant \text{ (с вер.} \geqslant 1-\delta, \text{ н-во Маркова) } \leqslant \frac{1}{\delta}\mathsf{E}_{(X,Y)\sim P}\bigl\{\mathsf{E}_\pi e^f\bigr\} \leqslant$$

$$\leqslant \text{ } (\pi \text{ неслучайно}) \text{ } \leqslant \frac{1}{\delta}\mathsf{E}_\pi\bigl\{\mathsf{E}_{(X,Y)\sim P} e^f\bigr\} \leqslant \text{(лемма Хевдинга)} \leqslant$$

$$\leqslant \frac{1}{\delta}\mathsf{E}_\pi e^{\lambda^2/(8\ell)} = \frac{1}{\delta} e^{\lambda^2/(8\ell)}$$

Первая теорема PAC-Bayes: доказательство

Вариационное определение KL-дивергенции:

$$\mathsf{E}_\rho f \leqslant \mathrm{KL}(\rho\|\pi) + \ln \mathsf{E}_\pi e^f. \tag{v}$$

Таким образом с учетом (v)

$$\mathsf{E}_{g\sim\rho}\big(R(g) - R_\ell(g)\big) \leqslant \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8\ell}.$$

Оптимизируя по λ, получим:

$$R(G_\rho) \leqslant R_\ell(G_\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2\ell}}.$$

■

Вспомогательная теорема PAC-Bayes

Всюду далее $\mathbb{R} = \{+1, -1\}$, $r(y_1, y_2) = [y_1 \neq y_2]$.

Теорема (Germain et. al. 2009)

Зафиксируем **любое** *априорное распределение* π *на множестве* $\mathcal{G}$ *с* $\pi(g) > 0$ *и выпуклую функцию* $D : [0,1] \times [0,1] \to \mathbb{R}$. *Тогда для всех* $\delta \in (0,1)$ *с вероятностью не меньше* $1 - \delta$ *(относительно случайного выпадения обучающей выборки) для* **всех** *апостериорных распределений* ρ **одновременно** *выплнено:*

$$D\bigl(R_\ell(G_\rho), R(G_\rho)\bigr) \leqslant$$
$$\leqslant \frac{1}{\ell}\left(\mathrm{KL}(\rho\|\pi) + \ln\left[\frac{1}{\delta}\mathsf{E}_{(X,Y)\sim P}\mathsf{E}_{g\sim\pi} e^{\ell D\bigl(R_\ell(G_\rho), R(G_\rho)\bigr)}\right]\right)$$

Вспомогательная теорема PAC-Bayes: доказательство

Н-во Маркова дает с вероятностью $\geqslant 1-\delta$:

$$\mathsf{E}_{g\sim\pi} e^{\ell D\left(R_\ell(G_\rho),R(G_\rho)\right)} \leqslant \frac{1}{\delta}\mathsf{E}_{(X,Y)\sim P}\mathsf{E}_{g\sim\pi} e^{\ell D\left(R_\ell(G_\rho),R(G_\rho)\right)}.$$

Логарифмируем обе части и перейдем к любому новому распределению ρ:

$$\ln\left[\mathsf{E}_{g\sim\rho}\frac{\pi(g)}{\rho(g)} e^{\ell D\left(R_\ell(G_\rho),R(G_\rho)\right)}\right] \leqslant \ln\left[\frac{1}{\delta}\mathsf{E}_{(X,Y)\sim P}\mathsf{E}_{g\sim\pi} e^{\ell D\left(R_\ell(G_\rho),R(G_\rho)\right)}\right].$$

Неравенство Йенсена дает:

$$\mathsf{E}_{g\sim\rho}\ln\left[\frac{\pi(g)}{\rho(g)} e^{\ell D\left(R_\ell(G_\rho),R(G_\rho)\right)}\right] \leqslant \ln\left[\frac{1}{\delta}\mathsf{E}_{(X,Y)\sim P}\mathsf{E}_{g\sim\pi} e^{\ell D\left(R_\ell(G_\rho),R(G_\rho)\right)}\right].$$

Вспомогательная теорема PAC-Bayes: доказательство

$$\mathsf{E}_{g\sim\rho} \ln\left[\frac{\pi(g)}{\rho(g)} e^{\ell D\left(R_\ell(G_\rho),R(G_\rho)\right)}\right] \leqslant \ln\left[\frac{1}{\delta}\mathsf{E}_{(X,Y)\sim P}\mathsf{E}_{g\sim\pi} e^{\ell D\left(R_\ell(G_\rho),R(G_\rho)\right)}\right].$$

Поскольку

$$\mathsf{E}_{g\sim\rho} \ln\left[\frac{\pi(g)}{\rho(g)}\right] = -\mathrm{KL}(\rho\|\pi),$$

а также в силу выпуклости функции D и

$$\mathsf{E}_{g\sim\rho}\left[\ell D\left(R_\ell(G_\rho), R(G_\rho)\right)\right] \geqslant \ell D\left(\mathsf{E}_{g\sim\rho} R_\ell(G_\rho), \mathsf{E}_{g\sim\rho} R(G_\rho)\right)$$

мы получаем

$$D\left(R_\ell(G_\rho), R(G_\rho)\right) \leqslant$$
$$\leqslant \frac{1}{\ell}\left(\mathrm{KL}(\rho\|\pi) + \ln\left[\frac{1}{\delta}\mathsf{E}_{(X,Y)\sim P}\mathsf{E}_{g\sim\pi} e^{\ell D\left(R_\ell(G_\rho),R(G_\rho)\right)}\right]\right) \quad \blacksquare$$

Вторая теорема PAC-Bayes

Следующая функция выпукла:

$$kl(q,p) \stackrel{\text{def}}{=} \mathrm{KL}([q,1-q]\|[p,1-p]) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}.$$

Теорема (Seeger, 2002)

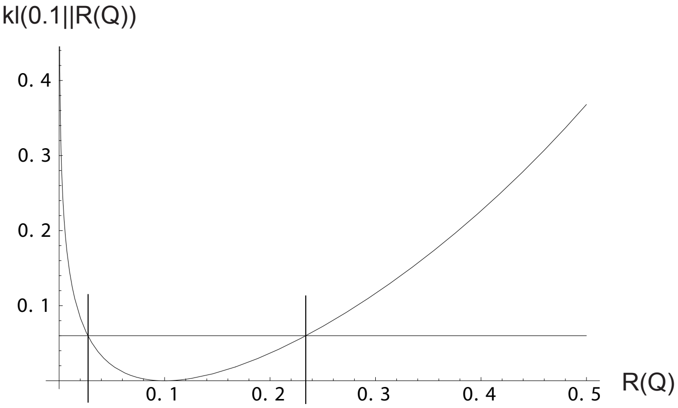
Зафиксируем **любое** *априорное распределение* π *на множестве* $\mathcal{G}$ *с* $\pi(g) > 0$. *Тогда для всех* $\delta \in (0,1)$ *с вероятностью не меньше* $1-\delta$ *(относительно случайного выпадения обучающей выборки) для* **всех** *апостериорных распределений* ρ **одновременно** *выплнено:*

$$kl\big(R_\ell(G_\rho), R(G_\rho)\big) \leqslant \frac{1}{\ell}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{\xi(\ell)}{\delta}\right],$$

где $\xi(\ell) = \sum_{i=0}^{\ell} C_\ell^i (i/\ell)^i (1-i/\ell)^{\ell-i} \leqslant \ell + 1$.

Вторая теорема PAC-Bayes: применение

Дает сразу и верхнюю и нижнюю оценки:



Вторая теорема PAC-Bayes: доказательство

Возьмем во вспомагательной теореме $D(q,p) = kl(q,p)$:

$$\mathsf{E}_{(X,Y)\sim P}\mathsf{E}_{g\sim\pi} e^{\ell kl\left(R_\ell(G_\rho), R(G_\rho)\right)} \leqslant$$

$$\leqslant \mathsf{E}_{g\sim\pi}\mathsf{E}_{(X,Y)\sim P}\left(\frac{R_\ell(g)}{R(g)}\right)^{\ell R_\ell(g)}\left(\frac{1-R_\ell(g)}{1-R(g)}\right)^{\ell\left(1-R_\ell(g)\right)} \leqslant$$

$$\leqslant \mathsf{E}_{g\sim\pi}\sum_{i=1}^{\ell}\underset{X^\ell\sim P^\ell}{P}\left(R_\ell(g)=\frac{i}{\ell}\right)\left(\frac{\frac{i}{\ell}}{R(g)}\right)^{i}\left(\frac{1-\frac{i}{\ell}}{1-R(g)}\right)^{\ell-i} =$$

$$= \mathsf{E}_{g\sim\pi}\sum_{i=1}^{\ell} C_\ell^i\left(\frac{i}{\ell}\right)^{i}\left(1-\frac{i}{\ell}\right)^{\ell-i} =$$

$$\sum_{i=1}^{\ell} C_\ell^i\left(\frac{i}{\ell}\right)^{i}\left(1-\frac{i}{\ell}\right)^{\ell-i}. \qquad \blacksquare$$

Первая теорема PAC-Bayes: revisited

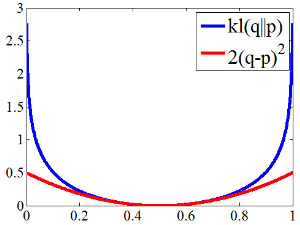
Неравенство Пинскера:

$$kl(q,p) \geqslant 2(q-p)^2.$$

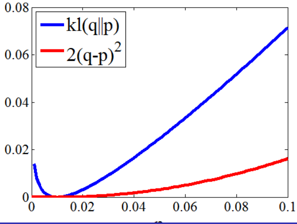
Отсюда из второй теоремы PAC-Bayes можно получить первую:

$$R(G_\rho) - R_\ell(G_\rho) \leqslant \frac{1}{2}\sqrt{kl\big(R_\ell(G_\rho), R(G_\rho)\big)} \leqslant \frac{1}{2}\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{\ell+1}{\delta}}{\ell}}.$$

$q = 0.5$



$q = 0.01$



PAC-Bayes vs. SLT

$$R(G_\rho) \leqslant R_\ell(G_\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2\ell}}.$$

Сходства:

- Количественные гарантии о качестве классификатора без ограничений вида P;
- Явная зависимость оценок от функции потерь.

Различия:

- Вместо «сложности класса» (например, VC-размерность) оценки учитывают сложность конкретного классификатора с помощью $\pi(g)$;
- Позволяет в явном виде включить в модель априор;
- Оценки получаются для классификатора Гиббса;
- Оценки остаются точными даже при $\mathrm{VCdim} = \infty$!

PAC-Bayes vs. SLT

$$R(G_\rho) \leqslant R_\ell(G_\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2\ell}}.$$

Сходства:

- Количественные гарантии о качестве классификатора без ограничений вида P;
- Явная зависимость оценок от функции потерь.

Различия:

- Вместо «сложности класса» (например, VC-размерность) оценки учитывают сложность конкретного классификатора с помощью $\pi(g)$;
- Позволяет в явном виде включить в модель априор;
- Оценки получаются для классификатора Гиббса;
- Оценки остаются точными даже при $\mathrm{VCdim} = \infty$!

PAC-Bayes vs. SLT

$$R(G_\rho) \leqslant R_\ell(G_\rho) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{2\ell}}.$$

Сходства:

- Количественные гарантии о качестве классификатора без ограничений вида P;
- Явная зависимость оценок от функции потерь.

Различия:

- Вместо «сложности класса» (например, VC-размерность) оценки учитывают сложность конкретного классификатора с помощью $\pi(g)$;
- Позволяет в явном виде включить в модель априор;
- Оценки получаются для классификатора Гиббса;
- Оценки остаются точными даже при $\mathrm{VCdim} = \infty$!

PAC-Bayes vs. Bayes

$$R(G_\rho) \leqslant R_\ell(G_\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2\ell}}.$$

Сходства:

- Учет априора в явном виде.

Различия:

- Количественные гарантии о качестве классификатора;
- Не нужны никакие предположения об адекватности априора;
- Оценки справедливы одновременно для всех апостериорных распределений ρ.

PAC-Bayes vs. Bayes

$$R(G_\rho) \leqslant R_\ell(G_\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2\ell}}.$$

Сходства:

- Учет априора в явном виде.

Различия:

- Количественные гарантии о качестве классификатора;
- Не нужны никакие предположения об адекватности априора;
- Оценки справедливы одновременно для всех апостериорных распределений ρ.

PAC-Bayes vs. Bayes

$$R(G_\rho) \leqslant R_\ell(G_\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2\ell}}.$$

Сходства:

- Учет априора в явном виде.

Различия:

- Количественные гарантии о качестве классификатора;
- Не нужны никакие предположения об адекватности априора;
- Оценки справедливы одновременно для всех апостериорных распределений ρ.

Применение PAC-Bayes оценок: линейные классификаторы

Рассмотрим линейные классификаторы вида

$$\mathcal{G} = \left\{ g_w(\mathbf{x}) = \operatorname{sgn}\left(\mathbf{w}^\top \varphi(\mathbf{x})\right) \right\},$$

проходящие через начало координат ($b = 0$).

- PAC-Bayes дает оценки для классификатора Гиббса.
- Как применить, например, для SVM?

Идея:

- Построить композицию классификаторов, эквивалентную нашему классификатору;
- Оценить риск этой композиции с помощью удвоенного риска Гиббса!

Применение PAC-Bayes оценок: линейные классификаторы

Положим
$$\rho = \mathcal{N}(\mu\mathbf{w}/\|\mathbf{w}\|, I), \quad \mu \in \mathbb{R}.$$

Тогда
$$\mathrm{sgn}\big(\mathbf{w}^\top\varphi(\mathbf{x})\big) = \mathrm{sgn}\big(\mathsf{E}_{w\sim\rho}g_w(\mathbf{x})\big);$$

и для соответствующего классификатора Гиббса:

$$R_\ell(G_\rho) = \frac{1}{\ell}\sum_{i=1}^{\ell}\tilde{F}\big(\mu\gamma(\mathbf{x}_i, y_i)\big);$$
$$\gamma(\mathbf{x}, y) = \frac{y\mathbf{w}^\top\varphi(\mathbf{x})}{\|\varphi(\mathbf{x})\|\,\|\mathbf{w}\|};$$
$$\tilde{F}(x) = 1 - \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx.$$

Применение PAC-Bayes оценок: линейные классификаторы

Теорема (Langford, Shawe-Taylor, 2002, 2005)

Для всех $\delta \in (0,1)$ с вероятностью не меньше $1-\delta$ (относительно случайной реализации обучающей выборки) **одновременно** *для всех классификаторов Гиббса G_ρ, $\rho = \mathcal{N}(\mu \mathbf{w}/\|\mathbf{w}\|, I)$, выполнено:*

$$kl\big(R_\ell(G_\rho)\|R(G_\rho)\big) \leqslant \frac{\frac{\mu^2}{2} + \ln\frac{\ell+1}{\delta}}{\ell}. \tag{1}$$

Рецепт:

- Настройте SVM и получите $\mathbf{w}$;
- Оптимизируйте оценку (1) по μ. Тогда

$$\mathsf{E}_{(\mathbf{x},y)\sim D}\left[\operatorname{sgn}(\mathbf{w}^\top \mathbf{x}) \neq y\right] \leqslant 2R(G_\rho).$$

Применение PAC-Bayes оценок: линейные классификаторы

(Langford, 2005)

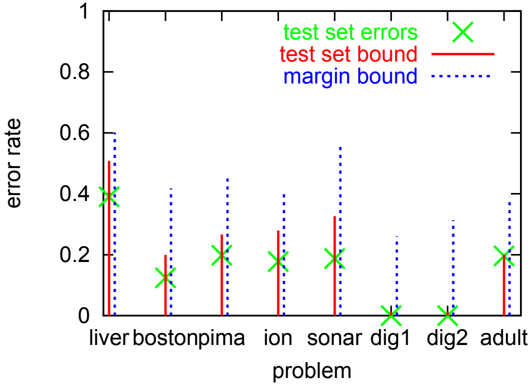


Figure 10: This figure shows the results of applying SVMlight to 8 datasets with a Gaussian kernel and a 70/30 train/test split. The observed test error rate is graphed as an X. On the test set, we calculate a binomial confidence interval (probability of bound failure equals 0.01) which upper bounds the true error rate. On the training set we calculate the PAC-Bayes margin bound for an optimized choice of μ.

Обучаем априорное распределение на данных

- Оценки зависят от KL-дивергенции между ρ и π;
- Чем лучше априор, тем лучше оценки;
- Давайте обучать априор по части обучающей выборки;
- Подставлять этот априор в оценку;
- Вычислять оценку по оставшимся данным.

Настроим SVM $\mathbf{w}_r$ по первым r объектам обучающей выборки.

Обозначим $\bar{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$.

Теперь введем априор $\pi = \mathcal{N}(\eta\,\bar{\mathbf{w}}_r, I)$.

Снова применим вторую PAC-Bayes теорему.

Обучаем априорное распределение на данных

Теорема (Ambroladze et. al., 2007)

Для всех $\delta \in (0,1)$ *с вероятностью не меньше* $1-\delta$ *(относительно случайной реализации обучающей выборки)* **одновременно** *для всех классификаторов Гиббса* G_ρ, $\rho = \mathcal{N}(\mu\bar{\mathbf{w}}, I)$, *выполнено:*

$$kl\big(R_{\ell-r}(G_\rho)\|R(G_\rho)\big) \leqslant \frac{\frac{\|\eta\bar{\mathbf{w}}_r - \mu\bar{\mathbf{w}}\|^2}{2} + \ln\frac{\ell-r+1}{\delta}}{\ell - r}, \qquad (1)$$

где

$$R_{\ell-r}(G_\rho) = \frac{1}{\ell-r}\sum_{i=r+1}^{\ell} \tilde{F}\big(\mu\gamma(\mathbf{x}_i, y_i)\big).$$

Важно: $\mathbf{w}$ можно вычислять по **всей** обучающей выборке!

Обучаем априорное распределение на данных: Еще лучше!

Введем J разных априоров $\pi_j = \mathcal{N}(\eta_j \mathbf{w}_r / \|\mathbf{w}_r\|, I)$.

Теорема (Ambroladze et. al., 2007)

Для всех $\delta \in (0,1)$ с вероятностью не меньше $1-\delta$ (относительно случайной реализации обучающей выборки) **одновременно** *для всех классификаторов Гиббса G_ρ, $\rho = \mathcal{N}(\mu\bar{\mathbf{w}}, I)$, и* **одновременно** *для всех j выполнено:*

$$kl\left(R_{\ell-r}(G_\rho) \| R(G_\rho)\right) \leqslant \frac{\frac{\|\eta_j \bar{\mathbf{w}}_r - \mu\bar{\mathbf{w}}\|^2}{2} + \ln\frac{\ell-r+1}{\delta} + \ln J}{\ell - r}.$$

- Для каждого j оптимизируем по μ;
- Выбираем лучшую из полученных оценок.

Обучаем априорное распределение на данных: SVM

Problem	# samples	input dim.	Pos/Neg
Handwritten-digits	5620	64	2791 / 2829
Waveform	5000	21	1647 / 3353
Pima	768	8	268 / 500
Ringnorm	7400	20	3664 / 3736
Spam	4601	57	1813 / 2788

Обучаем априорное распределение на данных

Problem		2FCV	10FCV	PAC	PrPAC
digits	Bound	–	–	0.175	0.107
	CE	0.007	0.007	0.007	0.014
waveform	Bound	–	–	0.203	0.185
	CE	0.090	0.086	0.084	0.088
pima	Bound	–	–	0.424	0.420
	CE	0.244	0.245	0.229	0.229
ringnorm	Bound	–	–	0.203	0.110
	CE	0.016	0.016	0.018	0.018
spam	Bound	–	–	0.254	0.198
	CE	0.066	0.063	0.067	0.077

Усреднение по 50 случайным разбиениям
на обучение/контроль (80%/20%).

Другие навправления...

- KL-дивергенцию можно заставить исчезнуть!
- Классификаторы, минимизирующие PAC-Bayes оценки;
- PAC-Bayes оценки плотности (непрерывной и дискретной);
- Приложения в collaborative filtering (co-clustering);
- "Non i.i.d", martingales, reinforcement learning;
- Приложения в tunsductive и unsupervised learning;
- Минимизируя PAC-Bayes оценки можно получить SVM, KL-Regularized AdaBoost, Kernel Ridge Regression,...;
- Distribution-dependant prior.

Обзор литературы

- *Amiran Ambroladze,Emilio Parrado-Hernandez, and John Shawe-Taylor*. **Tighter PAC-Bayesbounds**. *Advancesin Neural Information Processing Systems (NIPS)*, 2007.
- *Jean-Yves Audibert, Olivier Bousquet*. **Combining PAC-Bayesian and generic chaining bounds**. *Journal of Machine Learning Research*, 8:863–889, 2007.
- *Olivier Catoni*. **PAC-Bayesian supervised classifcation: The thermodynamics of statistical learning**. *IMS Lecture Notes Monograph Series*, 56, 2007.
- *John Shawe-Taylor, Yevgeny Seldin, Francois Laviolette*. **PAC-Bayesian Analysis in Supervised, Unsupervised, and Reinforcement Learning**. Tutorial, ECML-PKDD 2012, Bristol.

Обзор литературы

- *Pascal Germain, Alexandre Lacasse, Francois Laviolette, and Mario Marchand.* **PAC-Bayesian learning of linear classifers**. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.
- *John Langford and John Shawe-Taylor.* **PAC-Bayes and margins**. *Advances in Neural Information Processing Systems (NIPS)*, 2002.
- *John Langford.* **Tutorial on practical prediction theory for classification**. *Journal of Machine Learning Research*, 6:273–306, 2005.
- *Guy Lever, Francois Laviolette, and John Shawe-Taylor.* **Distribution-dependent PAC-Bayes priors**. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.

Обзор литературы

- *Matthias Seeger*. **PAC-Bayesian generalization error bounds for Gaussian process classification**. *Journal of Machine Learning Research*, 2002.
- *Yevgeny Seldin and Naftali Tishby*. **PAC-Bayesian analysis of co-clustering and beyond**. *Journal of Machine Learning Research*, 11, 2010.
  Yevgeny Seldin, Peter Auer, Francois Laviolette, John Shawe-Taylor, and Ronald Ortner. **PAC-Bayesian analysis of contextual bandits**. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.
- *Yevgeny Seldin,Francois Laviolette, Nicolo Cesa-Bianchi, John Shawe-Taylor, and Peter Auer*. **PAC-Bayesian inequalities for martingales**. *IEEE Transactions on Information Theory*, 2012.

Обзор литературы

- *John Shawe-Taylor and Robert C. Williamson*. **A PAC analysis of a Bayesian estimator**. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1997.
- *John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony*. **Structural risk minimization over data-dependent hierarchies**. *IEEE Transactions on Information Theory*, 44(5), 1998.
- *David McAllester*. **Some PAC-Bayesian theorems**. *Machine Learning*, 37, 1999.

videolectures.net:
Workshop on PAC Bayesian Learning, 2010, London.

Спасибо!

Спасибо за внимание!