Аддитивная регуляризация наивного линейного байесовского классификатора.

Шишковец Светлана Сергеевна

Московский физико-технический институт
Факультет управления и прикладной математики
Научный руководитель: д.ф.-м.н. К. В. Воронцов

Группа 274, 2016

Цель исследования

Наивный байсовский классификатор (NB):

- ⊖ предположение о независимости признаков
- ⊕ время обучения $O(n\ell)$ для n признаков, ℓ объектов
- ⊕ высокое качество классификации текстов
- ⊕ обладает линейной моделью классификации

Цель работы — обобщить NB:

- ослабить ограничение независимости путём регуляризации
- сохранить высокую скорость обучения
- сохранить линейность классификатора
- ввести отбор признаков

Задача классификации

Дано:
$X_i(x_1, ..., x_j)$ — объекты
$y_i \in \mathbb{Y} = \{c_0, c_1, ..., c_m\}$ — классы.
Найти:
Линейный наивный байесовский классификатор:

$$a(X) = \sum_{j=1}^{n} w_j x_j$$

Критерий: максимум регуляризованного правдоподобия

$$\sum_{i=1}^{l} \sum_{j=1}^{n} p(X_i(x_j)|\theta_j) + R(\theta) \to \max_{\theta},$$

где θ_j - параметр вероятностной модели, $R(\theta)$- регуляризатор

Экспоненциальное семейство плотностей

Рассмотрим **экспоненциальне семейство плотностей**

$$p(x|\theta) = \exp\left(\frac{x\theta - c(\theta)}{\varphi} + h(x,\varphi)\right),$$

где $c(\theta)$, $h(x,\varphi)$ — функциональные параметры распределения, θ —*сдвиг*,φ — *разброс*.

Теорема

Пусть одномерные плотности $p(x^j|\theta_y^j)$ принадлежат экспоненциальному семейству. Если $\Theta = (\theta_y^j)$ является точкой максимума правдоподобия, то

$$\theta_y^j = [c']^{-1}\left(\frac{1}{|X_y|}\sum_{x_i \in X_y} x_i\right).$$

Линейный наивный байесовский классификатор

Теорема

Пусть $\mathbb{Y} = \{-1, +1\}$, плотности $p(x^j|\theta_y^j)$ принадлежат экспоненциальному семейству плотностей и параметры разброса не зависят от класса, $\varphi_y^j = \varphi^j$. Тогда NB представляется в линейном виде

$$a(x) = sign\left(\sum_{j=1}^{n} x^j w_j - w_0\right),$$

причём

$$w_j = \frac{1}{\varphi^j} \sum_{y \in \mathbb{Y}} y\theta_y^j, \qquad j = 1, \ldots, n.$$

Аддитивная регуляризация

Дополнительные критерии качества $\mathcal{R}_q(w) \to \max$.
Взвешенная сумма с коэффициентами регуляризации τ_q:

$$\mathcal{R}(w) = \sum_{k=1}^{K} \tau_k \mathcal{R}_k(w) \to \max_w .$$

Задача максимизации регуляризованного правдоподобия :

$$\sum_{j=1}^{n} \sum_{y \in \mathbb{Y}} \sum_{x_i \in X_y} \left(\frac{x_i^j \theta_y^j - c(\theta_y^j)}{\varphi_y^j} \right) + \mathcal{R}(w(\Theta)) \to \max_{\Theta},$$

Регуляризованный линейный наивный байесовский классификатор

Теорема

Пусть $\mathbb{Y} = \{-1, +1\}$, плотности $p(x^j|\theta_y^j)$ принадлежат экспоненциальному семейству плотностей и параметры разброса не зависят от класса, $\varphi_y^j = \varphi^j$. Тогда точка максимума регуляризованного правдоподобия удовлетворяет системе уравнений

$$w_j = \frac{1}{\varphi^j} \sum_{y \in \mathbb{Y}} y[c']^{-1} \left(\frac{1}{|X_y|} \sum_{x_i \in X_y} x_i + \frac{y}{|X_y|} \frac{\partial \mathcal{R}}{\partial w_j} \right).$$

Регулярзаторы для отбора признаков

- Сжимающий регуляризатор по L_1-норме:

$$\mathcal{R}(w) = -\tau \sum_{j=1}^{n} |w_j|.$$

- Сжимающий регуляризатор по L_0-норме:

$$\mathcal{R}(w) = -\tau \sum_{j=1}^{n} [w_j > 0]$$

$||w||_0 = \#\{j = 1...|Y||w^j \neq 0\}$ -количество ненулевых весов для каждого класса.

Идея: удалять признаки с малым весом.

Многоклассовая классификация

Подход **«One-vs-All»** с «выделенным» классом $c_0 \in \mathbb{Y}$.
В задаче медецинской дифференциальной диагностики
c_0 — **абсолютно здоровые**.

Регуляризатор

Идея: как можно сильнее дистанцировать векторы весов различных классов $y, z \in \mathbb{Y}$ друг от друга.

$$R = -\sum_{j=1}^{n} \sum_{y>z} |w_{yj}||w_{zj}| \longrightarrow \max,$$

где w_{yj} - вес j-го признака для класса $y \in \mathbb{Y}$.

Прикладная задача

Дифференциальная диагностика заболеваний по ЭКГ методом В.М.Успенского.

Данные

ЭКГ, преобразованная в символьную последовательность и разбитая на триграммы (слова из трех букв). Признаки – частоты триграмм.

Всего 216 признаков, 18 болезней и класс абсолютно здоровых.

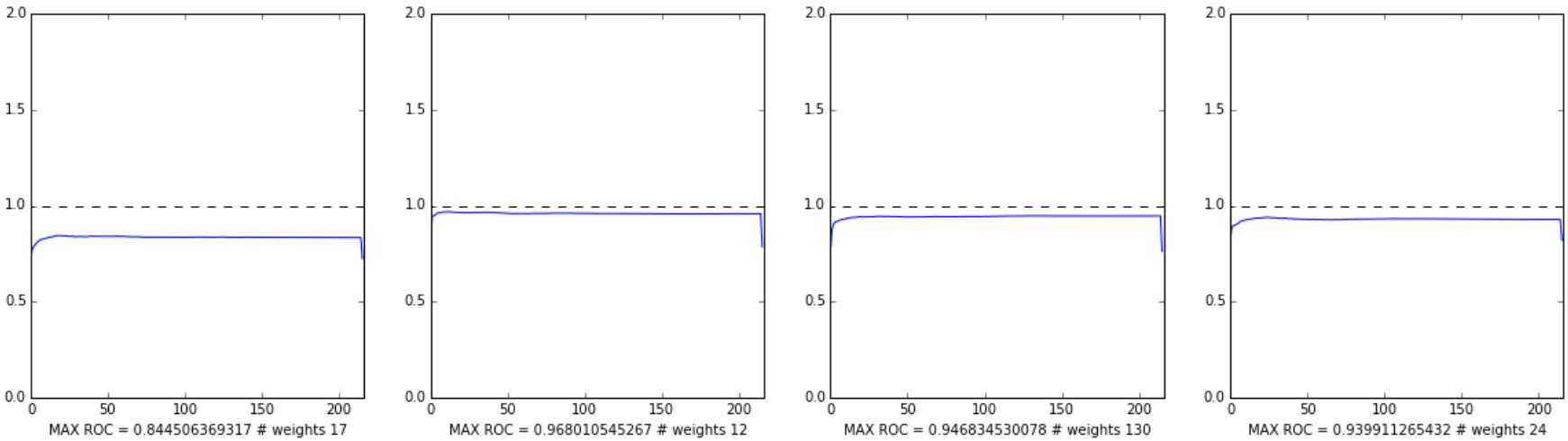
Цель

Повысить качество дифференциальной диагностики при помощи регуляризации линейного наивного байсовского классификатора.

В качестве критерия качества рассматривается площадь под ROC-кривой **AUC(Area Under Curve)**.

Отбор признаков

Зависимость значения *AUC* от количества используемых в классификаторе признаков для болезней: ВСД, НГБК, ГБ и ХГ1.



Как правило, в случае распределения Пуассона 10 – 20 информативных признаков достаточно. Однако, в случае нормального распределения необходимо рпссматривать 70 – 100 признаков.

Методы классификации

Выполним эксперемент для **нормального(Norm)** и **пуассоновского(Pois)** распределения. Для каждого исследуем отдельно ругуляризаторы отбора признаков (L_0, L_1) и их комбинацию с многоклассовым регуляризатором.

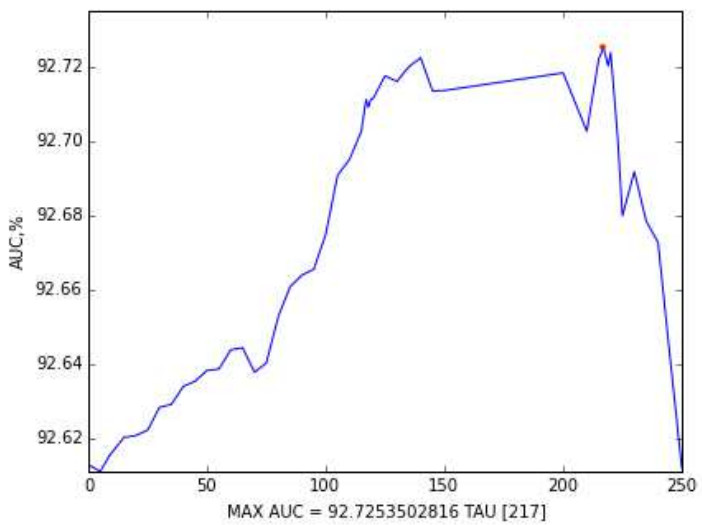
Результаты для распределения Пуассона

Болезнь	Pois	L_0	L_1	L_0+MR	L_1+MR
ВСД	0.8385	0.8252 (14)	**0.8401**	0.8245	0.8385
НГБК	0.9561	0.9510 (4)	0.9579	0.9516	**0.9593**
ГБ	0.9477	0.9466 (10)	0.9479	0.9466	**0.958**
ХГ1	0.9277	0.9202 (13)	**0.9298**	0.9178	0.9295
ДЖВП	0.9212	0.9189 (9)	0.9215	0.9182	**0.9259**
ЖКБ	0.9681	0.9662 (16)	0.9704	0.9670	**0.9798**
ИБС	0.9637	0.9613 (17)	0.9636	0.9615	**0.9699**
МКБ	0.9303	0.9272 (17)	0.9308	0.9274	**0.9386**
ММ	0.9124	0.9099 (12)	0.9130	0.9099	**0.9193**
ВСЕ	0.9268	0.9223 (11)	0.9280	0.9226	**0.9291**

Результаты для нормального распределения

Болезнь	Norm	L_0	L_1	L_0+MR	L_1+MR
ВСД	0.8435	0.8095 (91)	0.8448	0.8098	**0.8478**
НГБК	0.9588	0.9328 (49)	0.9594	0.9332	**0.9631**
ГБ	0.9395	0.9065 (117)	0.9397	0.9064	**0.9421**
ХГ1	0.9380	0.9138 (38)	0.9387	0.9146	**0.9412**
ДЖВП	0.9206	0.8895 (105)	0.9214	0.8896	**0.9201**
ЖКБ	0.9708	0.9590 (47)	0.9716	0.9586	**0.9791**
ИБС	0.9591	0.9397 (105)	0.9592	0.9393	**0.9620**
МКБ	0.9106	0.9010 (111)	**0.9278**	0.9010	0.9271
ММ	0.9428	0.8859 (99)	**0.9112**	0.8861	0.9106
ВСЕ	0.9268	0.9027 (88)	0.9273	0.9028	**0.9293**

Подбор коэффициентов регуляризации



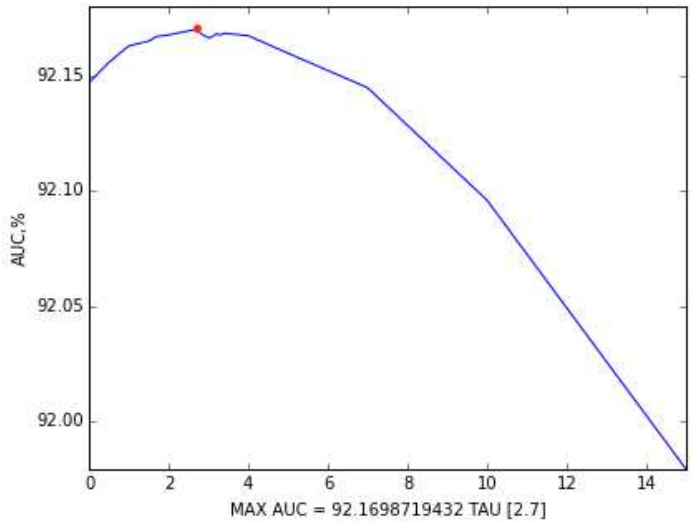


Рис.: Зависимость значения AUC от константы τ для L_1-регуляризатора(слева) и многоклассового регуляризатора(справа).

Выводы

По результатам проведенного эксперимента можно сделать следующие выводы:

- NB дает хорошие рузультаты как для нормального, так и для пуассоновского распределения
- обе регуляризации хорошо справляются с отбором признаков
- примененные типы регуляризаторов улучшают качество работы классификатора

Результаты, выносимые на защиту

- Предложен метод аддитивной регуляризации наивного байесовского классификатора
- Предложены регуляризаторы для отбора признаков и для повышения различности векторов весов признаков в многоклассовой классификации.
- Показано, что разработанные методы повышают качество дифференциальной диагностики заболеваний при использовании технологии информационного анализа электрокардиосигналов.