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:
$x_i = (x_i^1, ..., x_i^n)$ — объекты, $i = 1, ..., \ell$
$y_i \in \mathbb{Y} = \{c_0, c_1, ..., c_m\}$ — классы.
Найти:
Линейный наивный байесовский классификатор:

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

Критерий: максимум регуляризованного правдоподобия

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

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

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

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 $\mathcal{R}(w) \to \max$.
Взвешенная сумма с коэффициентами регуляризации τ_k:

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

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:

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

$||w||_0 = \#\left\{j = 1...|Y| \;\middle|\; w^j \neq 0\right\}$ — 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 медицинской 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. Признаки — частоты триграмм. 216 признаков, 18 болезней и класс абсолютно здоровых.

Цель:

Повысить качество дифференциальной диагностики при помощи регуляризации линейного NB. В качестве критерия качества рассматривается площадь под ROC-кривой **AUC(Area Under Curve)**.

Распределения:нормальное(Norm), Пуассона(Pois), Бернулли(Bern).

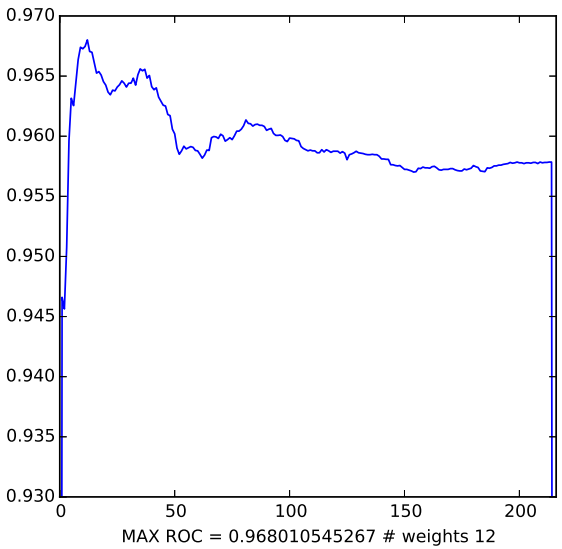
Для каждого исследуем регуляризаторы отбора признаков L_0, L_1 и их комбинацию с многоклассовым регуляризатором(MR).

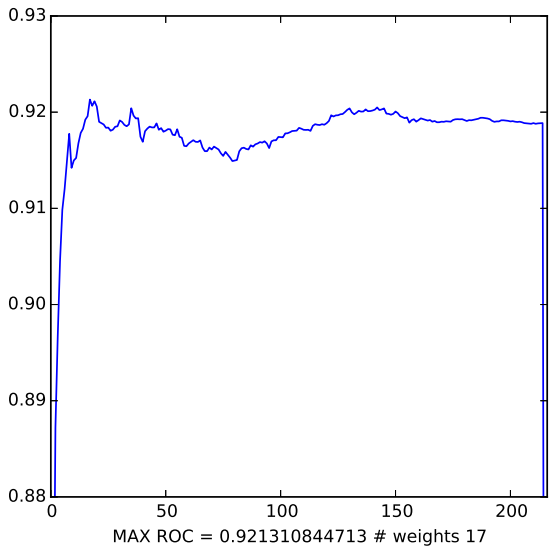
Объём исходных данных и аббревиатура

абсолютно здоровые	АЗ	193
аденома простаты	ДГПЖ	260
аднексит хронический	АХ	276
анемия железодефицитная	ЖДА	260
асептический некроз головки бедренной кости	НГБК	324
вегетососудистая дистония	ВСД	694
гипертоническая болезнь	ГБ	1894
дискинезия желчевыводящих путей	ДЖВП	717
желчнокаменная болезнь	ЖКБ	278
ишемическая болезнь сердца	ИБС	1265
миома матки	ММ	781
мочекаменная болезнь	МКБ	654
рак общий (онкопатология различной локализации)	РО	530
сахарный диабет (СД1 и СД2)	СД	871
узловой (диффузный) зоб щитовидный железы	УЩ	748
холецистит хронический	ХХ	340
хронический гастрит (гастродуоденит) гиперацидный	ХГ1	324
хронический гастрит (гастродуоденит) гипоацидный	ХГ2	700
язвенная болезнь	ЯБ	785

Отбор признаков

Зависимость *AUC* от количества признаков для ГБЭ(слева) и ЭА(справа) в случае Пуассона на обучающей выборке:





Для распределения Пуассона 10 информативных признака, для нормального 80 и для Бернулли 90.

Результаты для распределения Пуассона

Болезнь	Pois	L_0	L_1	L_0+MR	L_1+MR
ВДЭ	0.8340	0.8202(13)	**0.8401**(203)	0.8245(13)	0.8385(202)
ГБК	0.9578	0.9516(4)	0.9579(157)	0.9526(4)	**0.9593**(157)
ГБЭ	0.9464	0.9454(14)	0.9479(215)	0.9466(14)	**0.9580**(214)
ГДЭ	0.9280	0.9196(14)	**0.9298**(162)	0.9178(14)	0.9295(162)
ДЖЭ	0.9188	0.9157(7)	0.9215(201)	0.9182(7)	**0.9259**(201)
ЖКЭ	0.9682	0.9604(16)	0.9704(135)	0.9670(16)	**0.9798**(135)
ИБЭ	0.9626	0.9357(17)	0.9636(209)	0.9415(17)	**0.9699**(207)
МКЭ	0.9290	0.9272(17)	0.9308(194)	0.9279(17)	**0.9386**(193)
ММЭ	0.9120	0.9099(7)	0.9130(206)	0.9129(7)	**0.9193**(205)
РОЭ	0.9456	0.9425(12)	**0.9112**(181)	0.9561(12)	0.9106(180)
СДЭ	0.9524	0.9511(7)	0.9572(205)	0.9573(7)	**0.9604**(205)
УЩЭ	0.9318	0.9275(8)	0.9328(206)	0.9301(8)	**0.9335**(206)
ХГЭ	0.9323	0.9291(17)	0.9317(197)	0.9314(17)	**0.9342**(197)
ХХЭ	0.9324	0.9292(8)	0.9420(157)	0.9313(8)	**0.9496**(157)
ЭА	0.8696	0.8631(8)	**0.8802**(156)	0.8761(8)	0.8764(155)
ВСЕ	0.9261	0.9214 (11)	0.9280(154)	0.9286(11)	**0.9291**(153)

Результаты для нормального распределения

Болезнь	Norm	L_0	L_1	L_0+MR	L_1+MR
ВДЭ	0.8402	0.8290 (48)	0.8448(215)	0.80298(48)	**0.8478**
ГБК	0.9598	0.9454 (50)	0.9594(214)	0.9432(50)	**0.9631**(212)
ГБЭ	0.9389	0.9281 (112)	0.9397(216)	0.9264(112)	**0.9421**(215)
ГДЭ	0.9378	0.9323 (40)	0.9387(215)	0.9346(40)	**0.9412**(214)
ДЖЭ	0.9176	0.8993 (107)	**0.9214**(216)	0.8996(107)	0.9201(216)
ЖКЭ	0.9706	0.9696 (50)	0.9716(215)	0.9786(50)	**0.9791**(215)
ИБСЭ	0.9588	0.9464 (105)	0.9592 (216)	0.9493(105)	**0.9620**(216)
МКЭ	0.9267	0.9140 (108)	**0.9278** (216)	0.9210(108)	0.9271(215)
ММЭ	0.9103	0.9031 (115)	0.9112(216)	0.9161(115)	**0.9116**(215)
РОЭ	0.9418	0.9239 (106)	**0.9119**(216)	0.9261(105)	0.9209(216)
СДЭ	0.9503	0.9455 (89)	0.9513(214)	0.9471(88)	**0.9516**(214)
УЩЭ	0.9294	0.9079 (110)	0.9302(215)	0.9091(110)	**0.9304**(215)
ХГЭ	0.9296	0.9133 (115)	0.9315(216)	0.9161(115)	**0.9357**(214)
ХХЭ	0.9310	0.9180 (97)	**0.9323**(216)	0.9172(97)	0.9210(216)
ЭА	0.8766	0.8713 (44)	0.8812(216)	0.8861(44)	**0.8813** (215)
ВСЕ	0.9262	0.9145 (80)	0.9273(215)	0.9258(80)	**0.9303**(214)

Результаты для распределения Бернулли

Болезнь	Norm	L_0	L_1	L_0+MR	L_1+MR
ВДЭ	0.8316	0.8001 (123)	0.8348(216)	0.8079(123)	**0.8431**(215)
ГБК	0.8471	0.8354 (99)	0.8540(215)	0.8411(99)	**0.8588**(214)
ГБЭ	0.8653	0.8602 (117)	0.8674(216)	0.8621(117)	**0.8701**(216)
ГДЭ	0.8804	0.8782 (86)	0.8837(215)	0.8808(85)	**0.8842**(215)
ДЖЭ	0.8910	0.8895 (105)	**0.8844**(214)	0.8921(105)	0.8820(214)
ЖКЭ	0.9006	0.8990 (111)	**0.9178**(216)	0.9010(111)	0.9171(216)
ИБСЭ	0.8521	0.8399 (105)	0.8592(215)	0.8393(105)	**0.8607**(214)
МКЭ	0.9106	0.9010 (111)	**0.9278**(215)	0.9010(111)	0.9271(215)
ММЭ	0.9028	0.8959 (107)	**0.9108**(214)	0.8961(107)	0.9093(214)
РОЭ	0.9428	0.8859 (72)	0.9112(216)	0.8861(72)	**0.9106**(216)
СДЭ	0.8421	0.8390 (100)	0.8462(215)	0.8408(100)	**0.8479**(215)
УЩЭ	0.8807	0.8791 (78)	0.8819(214)	0.8825(78)	**0.8898**(214)
ХГЭ	0.8646	0.8498 (107)	0.8698(213)	0.8481(107)	**0.8719**(213)
ХХЭ	0.9021	0.8991 (99)	**0.9157**(215)	0.9053(99)	0.9134(215)
ЭА	0.9238	0.9159 (96)	0.9292(215)	0.9201(96)	**0.9309**(215)
ВСЕ	0.8825	0.8749 (95)	0.8893(215)	0.8808(95)	**0.8903**(215)

Подбор коэффициентов регуляризации

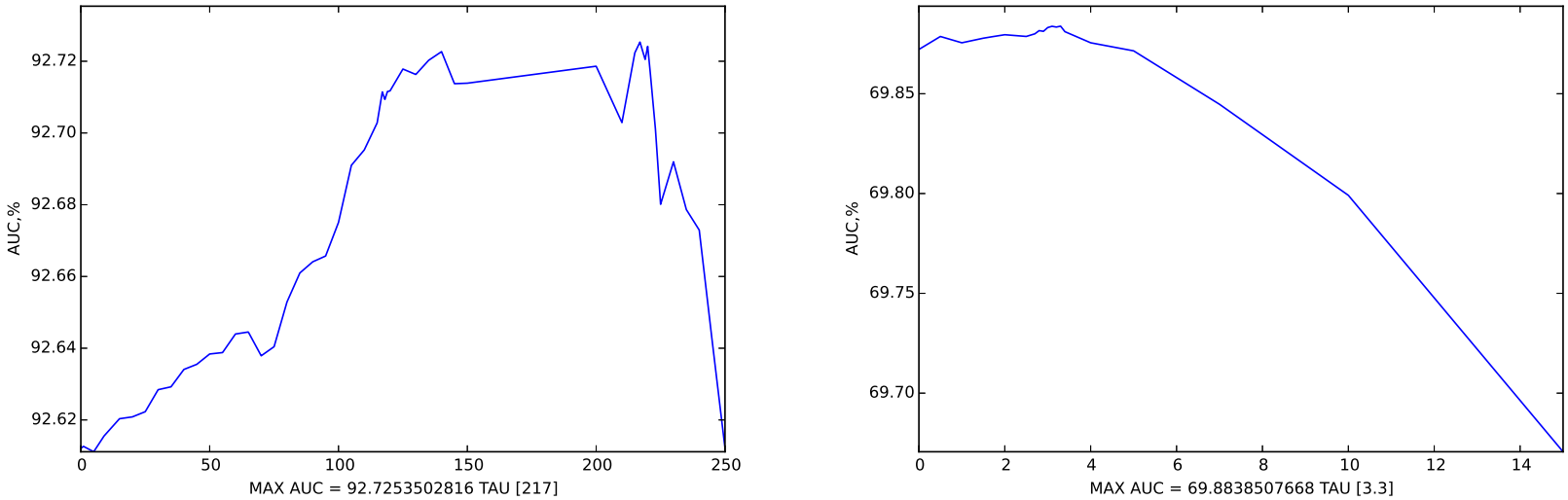


Рис.: Зависимость значения AUC от константы τ для L_1-регуляризатора(слева) и многоклассового регуляризатора(справа).

Выводы

По результатам проведенного эксперимента можно сделать следующие выводы:

- NB дает хорошие результаты как для нормального, так и для пуассоновского распределения
- результаты, полученные для распределения Бернулли значительно хуже
- 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.
- Выявлено ,что для распределения Пуассона при дифференциальной медицинской диагностике достаточно учитывать только 5% признаков.
- Показано, что разработанные методы повышают качество дифференциальной диагностики заболеваний при использовании технологии информационного анализа электрокардиосигналов.