Применение регуляризованной байесовской логистической регрессии в задаче кредитного скоринга

Боровых Н.И., Красоткина О.В.

Тульский государственный университет

ИОИ-2014
Крит, 4 - 11 октября 2014

План

Задача кредитного скоринга

Задача

Задача **кредитного скоринга** - это задача оценивания шансов соискателя на получение кредита на основе имеющейся информации о нем (возраст, стаж, средний доход семьи и т.д.). Необходимо произвести классификацию заемщиков на два класса - тех, кто сможет и не сможет выплатить кредит.

Предмет исследования

- Выделение информативных признаков среди всего множества признаков.
- Предсказание благонадежности клиента по известному вектору признаков.

Модели и методы, применяемые в задаче кредитного скоринга

- Линейный дискриминантный анализ
- Логистическая регрессия
- Деревья решений
- Решающий список логических правил
- Экспертные системы

Логистическая регрессия

[Dorte K., Tue Tjur (1999). Credit scoring: Discussion of methods and a case study, Department of Management Science and Statistics Copenhagen Business School Denmark] Сравниваются дискриминантный анализ, логистическая регрессия и нейронная сеть. Продемонстрирована эффективность логистической регрессии по сравнению с другими моделями.
[Komorad, K. (2002). On Credit Scoring Estimation. Master's Thesis. Humboldt University, Berlin.] Сравниваются логистическая регрессия, многослойный перцептрон и сеть радиальных базисных функций. Было показано, ошибка классификации при использовании логистической регрессии самая низкая.
[Thomas, L.C. (2009). Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press, Oxford.] Признает логистическую регрессию одним из самых широко используемых и эффективных методов.

Математическая постановка задачи

Пусть каждый кандидат $\omega \in \Omega$ описывается вектором числовых признаков $x = (x_1(\omega)), ..., x_n(\omega)) \in X = R^n$ и характеризуется некой скрытой переменной $y \in Y$, принимающей значения из множества $Y \in \{-1, +1\}$, где класс с меткой -1 соответствует классу «плохих»заемщиков, а +1 - «хороших».
Имеется обучающая выборка $D = \{\mathbf{x}_i, y_i\}_{i=1}^m$.
Решение задачи классификации удобно получать в виде оценок условных вероятностей принадлежности объекта к каждому из классов. Тогда, очевидно, решение можно найти:

$$\hat{y}_{MAP}(x) = \arg\max_c p(y = c|\mathbf{x}, \mathbf{w})$$

Модель логистической регрессии позволяет получить решение как раз в таком виде.

Модель логистической регрессии

Условное распределение зависимой переменной в случае двух классов представляет собой распределение Бернулли:

$$p(y|\mathbf{x}, \mathbf{w}) = Ber(y|sigm(\mathbf{w}^T\mathbf{x}))$$

где $sigm(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$ - логистическая функция
$\mathbf{w}^T\mathbf{x} =$ - линейная гиперплоскость разделяющая классы

$$p(y = -1|x, w) = \frac{1}{1 + exp(-\mathbf{w}^T\mathbf{x})}$$

$$p(y = +1|x, w) = \frac{1}{1 + exp(+\mathbf{w}^T\mathbf{x})}$$

Способы отбора признаковой информации

Методы-фильтры

Фильтры применяются на множестве всех признаков до восстановления зависимости, независимо от используемого метода восстановления. Как правило используются различные переборные стратегии. Метод отбора признаков не учитывает особенности искомой зависимости и используемых для ее восстановления алгоритмов.

Встроенные методы

Встроенные методы отбора признаков непосредственно инкорпорируются в метод решения задачи и, следовательно, существенно зависят от его специфики. Окончание процесса обучения одновременно является окончанием процесса отбора признаков.

Байесовский подход к отбору признаков - регуляризация

$$p(\mathbf{w}|D) \propto \underbrace{\left[\prod_{j=1}^{m} p(y_j \mid \mathbf{x}_j, \mathbf{w})\right]}_{p(D|\mathbf{w})} p(\mathbf{w}) \to \max_{\mathbf{w}}$$

$\mathbf{w}$ - вектор параметров решающего правила
$p(D|\mathbf{w})$ - совместная плотность распределения наблюдений в обучающей выборке, в предположении, что наблюдения независимы
$p(\mathbf{w})$ - априорная плотность распределения вектора параметров. При различных предположениях о виде априорного распределения получаем различные методы регуляризации.

Параметрическое семейство априорных плотностей распределения вектора параметров

Гребневая регрессия (L_2 - регуляризация). Следует из предположения, что вектор параметров - случайный вектор, распределенный по нормальному закону

$p(\mathbf{w}, \rho) = \prod_{i=1}^{p} p(w_i|\rho) \propto \exp(-\rho \sum_{i=1}^{p} w_i^2)$
Hoerl, A. E. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12: 55-67

Lasso (L_1 - регуляризация). Следует из предположения, что вектор параметров - случайный вектор, распределенный по закону Лапласса

$p(\mathbf{w}|\rho) = \prod_{i=1}^{p} p(w_i|\rho) \propto \exp(-\rho \sum_{i=1}^{p} |w_i|)$ Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B 58: 267-288

Параметрическое семейство априорных плотностей распределения вектора параметров

Elastic Net (линейная комбинация L_1 и L_2 штрафов)

$p(\mathbf{w}|\lambda_1, \lambda_2) = \prod_{i=1}^{p} p(w_i|\lambda_1, \lambda_2) \propto \exp\left[-\sum_{i=1}^{p} (\lambda_1|w_i| + \lambda_2 w_i^2)\right]$

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society Series B. 67(2): 301-320.

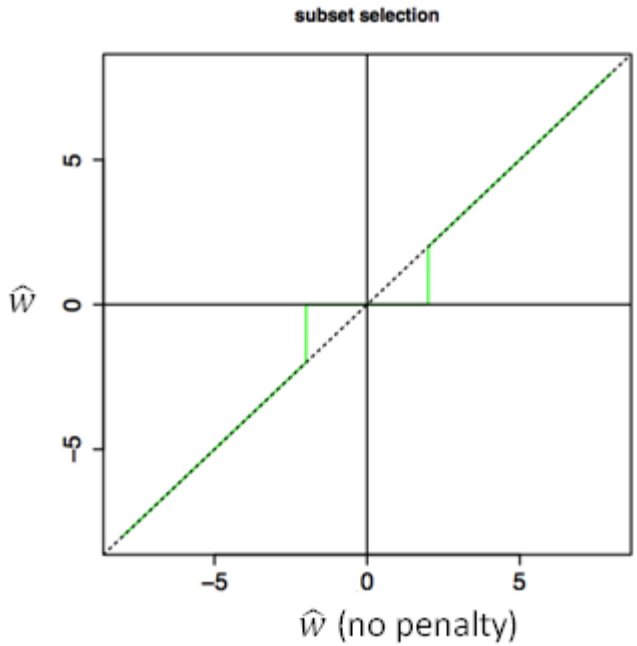
Свойства регуляризованных оценок

- **селективность** - способность исключать из модели незначащие признаки
- **несмещенность** - для отобранных в модель признаков априорное распределение коэффициентов регрессии является равномерным.
- **способность к отбору коррелированных регрессоров** свойство его оставлять в итоговой модели значащие признаки даже в случае наличия корреляции между ними
- **выполнение асимптотических оракульных неравенств**, оценивающие квадратичную сходимость абсолютных значениях регрессионных коэффициентов к их истинным значениям
- **верхняя граница риска оценки вектора регрессионных коэффициентов** не должна быть бесконечной величиной
- **непрерывность** - оценивание параметров модели возможно для всех значений структурного параметра и во всех точках признакового пространства

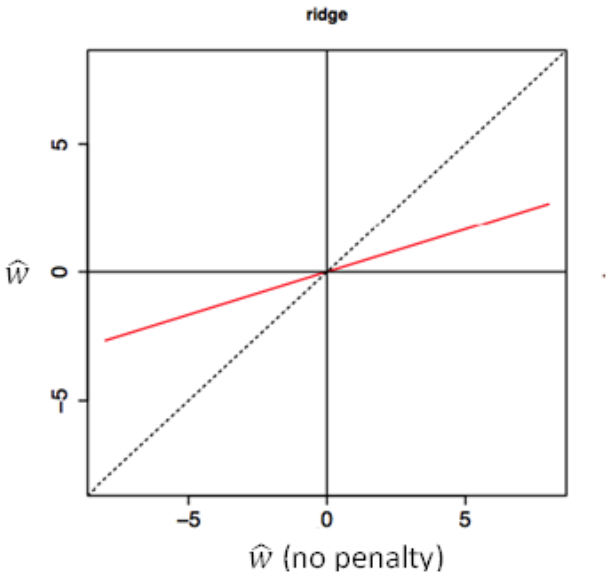
Свойства регуляризованных оценок

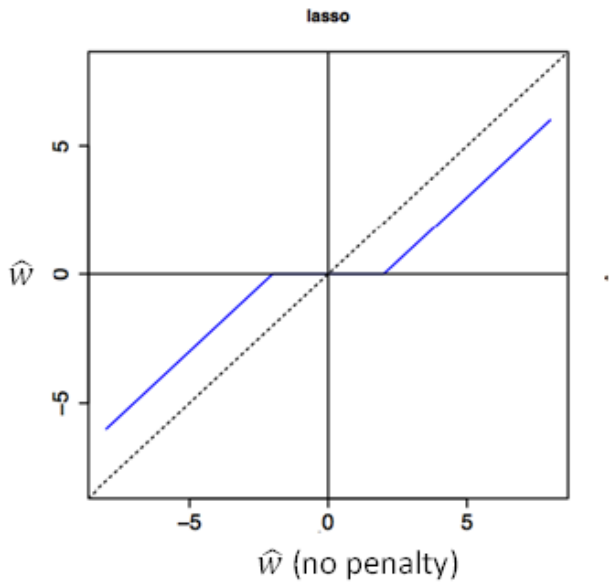
	Ridge	LASSO	ENet
Селективность	$-$	$+$	$+$
Несмещенность	$-$	$-$	$-$
Отбор коррелированных регрессоров	$+$	$-$	$+$
Конечность верхней границы риска	$+$	$-$	$-$
Выполнение оракульных неравенств	$-$	$-$	$-$
Непрерывность	$+$	$-$	$-$

Несмещенность, ассимптотическая несмещенность (unbiasedness) с точки зрения байесовского подхода выражается в том, для что для отобранных в модель признаков априорное распределение коэффициентов регрессии должно быть равномерным, т.е. $p'_\rho(|w_i|) \to 0$, когда $|w_i| \to \infty$



Несмещенность, ассимптотическая несмещенность (unbiasedness) с точки зрения байесовского подхода выражается в том, для что для отобранных в модель признаков априорное распределение коэффициентов регрессии должно быть равномерным, т.е. $p'_\rho(|w_i|) \to 0$, когда $|w_i| \to \infty$





Выберем в качестве априорной плотности распределения компонент w_i нормальное распределение с нулевыми математическим ожиданием и некоторой дисперсией r_i

$$p(w_i|r_i) \propto \left(1/r_i\right)^{-1/2} \exp\left(-w_i^2/(2r_i)\right).$$

Тогда, совместная априорная плотности распределения вектора параметров $\mathbf{w}$ примет вид

$$p(w_1,\ldots,w_n|r_1,\ldots,r_n) \propto \prod_{i=1}^{n}(1/r_i)^{-1/2} \exp\left(-\sum_{i=1}^{n}(w_i^2/(2r_i)\right).$$

Пусть величины обратные дисперсиям имеют гамма-распределение

$$p\big((1/r_i)|\alpha,\beta\big) \propto (1/r_i^{\alpha-1})\exp\big(-\beta(1/r_i)\big).$$

Тогда, совместная априорная плотности распределения дисперсий $1/r_i$ примет вид

$$p(1/r_1,\ldots,1/r_n|\alpha,\beta) \propto \Big(\prod_{i=1}^{n}(1/r_i^{\alpha-1})\exp\big(-\beta(1/r_i)\Big).$$

Для наделения критерия свойством отбрасывания нерелевантных признаков, выберем параметры гамма распределения следующим образом $\alpha = 1 + 1/(2\mu), \beta = 1/(2\mu)$,

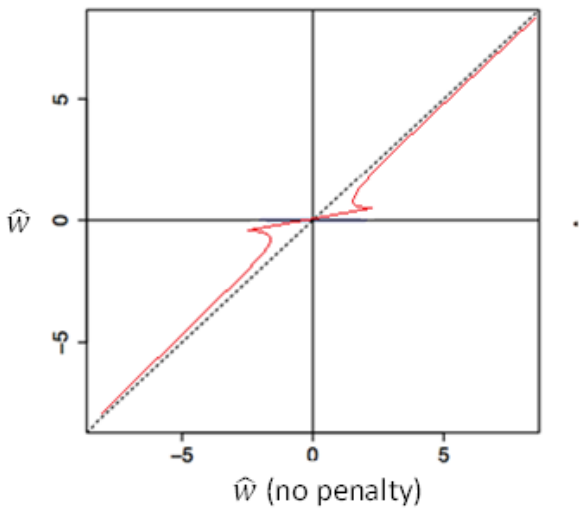
Свойства регуляризованных оценок

Селективность

$$\mu \to 0$$
$$\begin{cases} E\left(1/r_i\right) \to 1 \\ Var\left(1/r_i\right) \to 0 \\ \left(\sqrt{Var\left(1/r_i\right)} \Big/ E\left(1/r_i\right)\right) \to 0 \end{cases}$$

$$\mu \to \infty$$
$$\begin{cases} E\left(1/r_i\right) \to \infty \\ Var\left(1/r_i\right) \to \infty \\ \left(\sqrt{Var\left(1/r_i\right)} \Big/ E\left(1/r_i\right)\right) \to 1 \end{cases}$$

Несмещенность, ассимптотическая несмещенность (unbiasedness) с точки зрения байесовского подхода выражается в том, для что для отобранных в модель признаков априорное распределение коэффициентов регрессии должно быть равномерным, т.е. $p'_\mu(|w_i|) \to 0$, когда $|w_i| \to \infty$



Принцип максимизации совместной апостериорной плотности

$$p(\mathbf{w}|D) \propto p(D|\mathbf{w})p(\mathbf{w}|\mathbf{r})p(\mathbf{r}|\mu) \to \max_{\mathbf{w},\mathbf{r}}$$

приводит к следующему критерию обучения

$$J(\mathbf{w},\mathbf{r}) = -\sum_{i=1}^{m} \ln(sigm(y_i \sum_{j=1}^{n} w_j x_{ij})) + \frac{1}{2}\sum_{i=1}^{n} \frac{w_i^2}{r_i} + (\alpha - \frac{1}{2})\sum_{i=1}^{n} \ln r_i + \beta \sum_{i=1}^{n} \frac{1}{r_i} \to \min_{\mathbf{w},\mathbf{r}} \quad (1)$$

Будем минимизировать данный критерий методом Гаусса-Зайделя по двум группам переменных $(\mathbf{w}, \mathbf{r})$.
Пусть $\left(\mathbf{w}^{(\mathbf{k})}, \mathbf{r}^{(\mathbf{k})}\right)$ — очередное приближение к точке минимума.

Следующее значение вектора дисперсий можно получить, приравняв к нулю частные производные критерия максимального правдоподобия по каждой компоненте вектора

$$r_i^{(k+1)} = \left(\mu (w_i^{(k)})^2 + 1\right) / \left(\mu + 1\right).$$

Для получения очередного значения вектора коэффициентов $\mathbf{w}^{(\mathbf{k}+\mathbf{1})}$ решим оптимизационную задачу с помощью метода Ньютона

$$J(\mathbf{w}) = -\sum_{i=1}^{m} \ln(sigm(y_i \sum_{j=1}^{n} w_j x_{ij})) + \frac{1}{2}\sum_{i=1}^{n} \frac{w_i^2}{r_i} \to min(\mathbf{w})$$

Экспериментальное исследование на модельных данных

- Пусть каждое значение вектора признаков распределено независимо по нормальному закону $x_{i,j}: N(0,1), i = 1, ..., m, j = 1, ..., n$.
- Пусть число наблюдений n=50, число признаков 100
- Во всех экспериментах только 2 признака были значащими
- В ходе экспериментов предложенный подход сравнивался с методами восстановления регуляризованной логистической регрессии Ridge, Lasso и ElasticNet.
- в таблице приведена доля ошибочно классифицированных объектов на контрольной совокупности

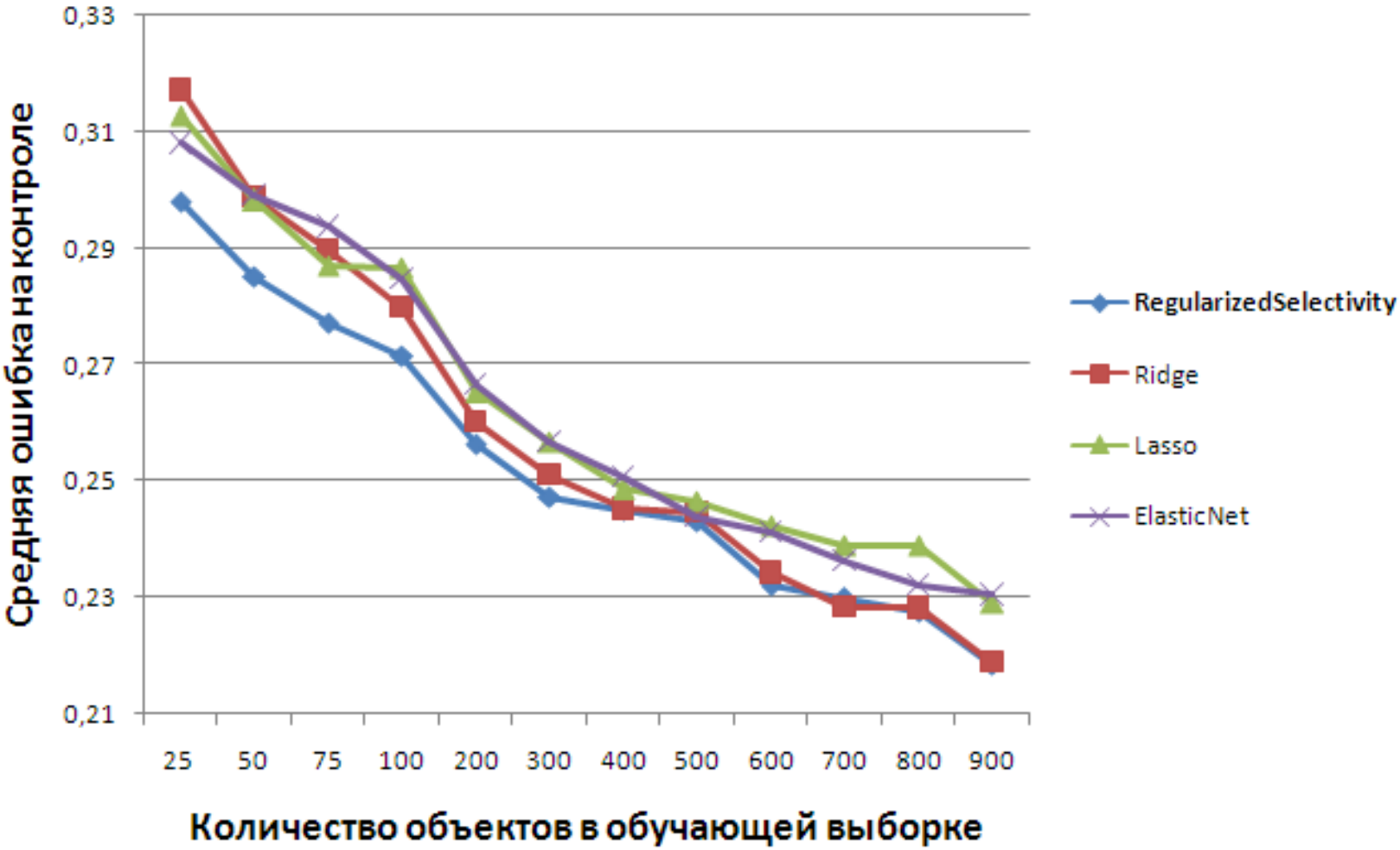
Ridge	Lasso	Elastic Net	**SupervisedSel**
0.389	0.117	0.14	**0.095**

Экспериментальное исследование на реальных данных
Описание данных

- В качестве данных для исследования взяты данные German Credit Data из репозитория UCI . https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)
- Данные содержат 24 признака, 1000 объектов.
- Для иллюстрации влияния зависимости соотношения количества объектов в выборке и количества признаков на качество распознавания эксперименты проводились для различного числа элементов в обучающей выборке. Для каждого значения проводилось 20 экспериментов, по которым показатели качества усреднялись.

Экспериментальное исследование на реальных данных

Результаты экспериментов



Выводы

- Исследовано применение модели регуляризованной логистической регрессии в задаче банковского скоринга
- Рассмотрены свойства получаемых оценок для случая гребневой регрессии, регуляризации Lasso и Elastic net
- Предложен новый метод регуляризации для задачи логистической регрессии, позволяющий отбирать параметры модели при соблюдении ассимптотической несмещенности оценок
- Проведено сравнительное исследование предложенного подхода на модельных и реальных данных, показывающее его эффективность