Введение в Data Science
Занятие 1. Классификация и регрессия

Николай Анохин Михаил Фирулик

4 марта 2014 г.

Постановка задач классификации и регрессии

Теория принятия решений

Обучение модели

Выбор модели

Классификация: интуиция

Задача

Разработать алгоритм, позволяющий определить класс произвольного объекта из некоторго множества

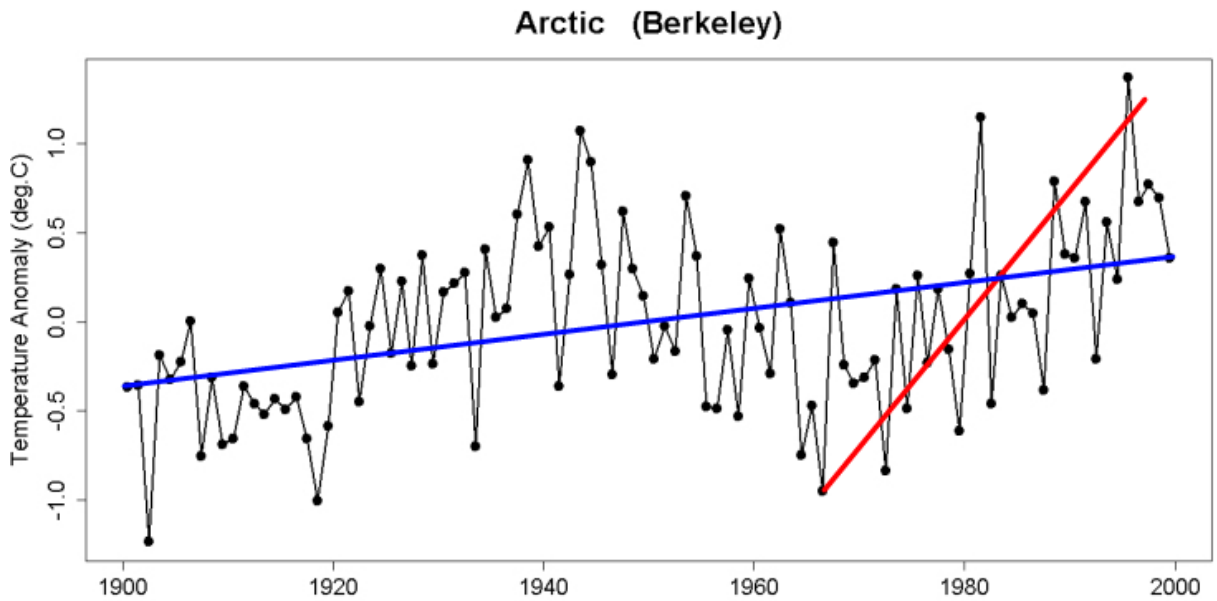
- Дана *обучающая выборка*, в которой для каждого объекта известен класс

Регрессия: интуиция

Задача

Разработать алгоритм, позволяющий предсказать числовую характеристику произвольного объекта из некоторого множества

- Дана *обучающая выборка*, в которой для каждого объекта известно значение данной числовой характеристики



Формализуем

X – множество объектов
T – множество значений целевой переменной (target variable)

Дана обучающая выборка из объектов

$$\mathbf{X} = (x_1, \ldots, x_N)^\top, \; x_i \in X$$

и соответствующие им классы

$$\mathbf{T} = (t_1, \ldots, t_N)^\top, \; t_i \in T$$

Требуется найти функцию

$$y^*(x) : X \to T,$$

позволяющую для произвольного $x \in X$ наиболее точно предсказать соответствующее $t \in T$

Целевая переменная

- $T = \{C_1, \ldots, C_K\}$ – задача классификации в K непересекающихся классов
- $T = [a, b] \subset R$ – задача регрессии

Как решать?

M Выдвигаем гипотезу насчет **модели** - семейства параметрических функций вида

$$Y = \{y(x, \theta) : X \times \Theta \to T\},$$

которая могла бы решить нашу задачу (model selection)

L Выбираем наилучшие параметры модели θ^*, используя **алгоритм обучения**

$$A(\mathbf{X}, \mathbf{T}) : (X, T)^N \to Y$$

(learning/inference)

D Используя полученную модель $y^*(x) = y(x, \theta^*)$, классифицируем неизвестные объекты (decision making)

Теория принятия решений

M Выдвигаем гипотезу насчет **модели** - семейства параметрических функций вида

$$Y = \{y(x, \theta) : X \times \Theta \to T\},$$

которая могла бы решить нашу задачу (model selection)

L Выбираем наилучшие параметры модели θ^*, используя **алгоритм обучения**

$$A(\mathbf{X}, \mathbf{T}) : (X, T)^N \to Y$$

(learning/inference)

D Используя полученную модель $y^*(x) = y(x, \theta^*)$, классифицируем неизвестные объекты (decision making)

Что моделировать?

Генеративные модели. Смоделировать $p(x|C_k)$ и $p(C_k)$, применить теорему Байеса

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)}$$

и использовать $p(C_k|x)$ для принятия решения
(NB, Bayes Networks, MRF)

Дискриминативные модели. Смоделировать $p(C_k|x)$ и использовать ее для принятия решения
(Logistic Regression, Decision Trees)

Функции решения. Смоделировать напрямую $f(x) : X \to T$
(Linear Models, Neural Networks)

Минимизируем риск

Пусть
$\mathcal{R}_k$ – область, такая что все $x \in \mathcal{R}_k$ относим к C_k
Дано
R_{kj} – риск, связанный с отнесением объекта класса C_k к классу C_j
Найти
$\forall k : \mathcal{R}_k$, такие, что математическое ожидание риска $E[R]$ минимально.

$$E[R] = \sum_k \sum_j \int_{\mathcal{R}_j} R_{kj} p(C_k|x) p(x) dx$$

Медицинская диагностика

Матрица риска $[R_{kj}]$

	sick	normal
sick	0	10
normal	1	0

Условные вероятности $p(C_k|x)$

$$p(\texttt{normal}|\texttt{moving}) = 0.9,\ p(\texttt{normal}|\texttt{not moving}) = 0.3$$

Вероятности $p(x)$

$$p(\texttt{moving}) = 0.7$$

Требуется определить $\mathcal{R}_{\texttt{sick}}$, $\mathcal{R}_{\texttt{normal}}$

Регрессия

Те же виды моделей: **генеративные**, **дискриминативные**, **функция решения**

Задана функция риска

$$R(t, y(x))$$

Математическое ожидание $E[R]$

$$E[R] = \int\int R(t, y(x))p(x, t)dxdt$$

Для квадратичной функции риска $R(t, y(x)) = [t - y(x)]^2$

$$y(x) = E_t[t|x]$$

Когда удобнее вероятностные модели

- Функция риска может меняться
- Отказ от классификации (reject option)
- Дисбаланс в выборке
- Ансамбли моделей

Обучение модели

M Выдвигаем гипотезу насчет **модели** - семейства параметрических функций вида

$$Y = \{y(x, \theta) : X \times \Theta \to T\},$$

которая могла бы решить нашу задачу (model selection)

L Выбираем наилучшие параметры модели θ^*, используя **алгоритм обучения**

$$A(\mathbf{X}, \mathbf{T}) : (X, T)^N \to Y$$

(learning/inference)

D Используя полученную модель $y^*(x) = y(x, \theta^*)$, классифицируем неизвестные объекты (decision making)

Выбор параметров модели

Функция потерь $\mathcal{L}(x, t, \theta)$ - ошибка, которую для данного x дает модель $y(x, \theta)$ по сравнению с реальным значением t

Эмпирический риск – средняя ошибка на обучающей выборке

$$Q(\mathbf{X}, \mathbf{T}, \theta) = \frac{1}{N} \sum_{k=1}^{N} \mathcal{L}(x_k, t_k, \theta)$$

Задача – найти значение θ^*, минимизирующее эмпирический риск

$$\theta^* = \theta^*(\mathbf{X}, \mathbf{T}) = \text{argmin}_\theta Q(\mathbf{X}, \mathbf{T}, \theta)$$

Некоторые функции потерь

- Индикатор ошибки

$$\mathcal{L}(x, t, \theta) = 0 \text{ if } y(x, \theta) = t \text{ else } 1$$

- Функция Минковского

$$\mathcal{L}(x, t, \theta) = |t - y(x, \theta)|^q$$

Частные случаи: квадратичная $q = 2$, абсолютная ошибка $q = 1$

- Hinge

$$\mathcal{L}(x, t, \theta) = \max(0, 1 - t * y(x, \theta))$$

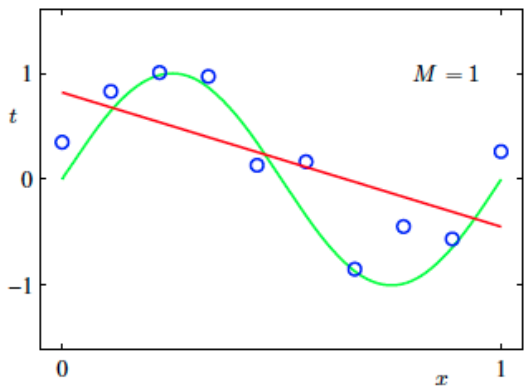
- Информационная

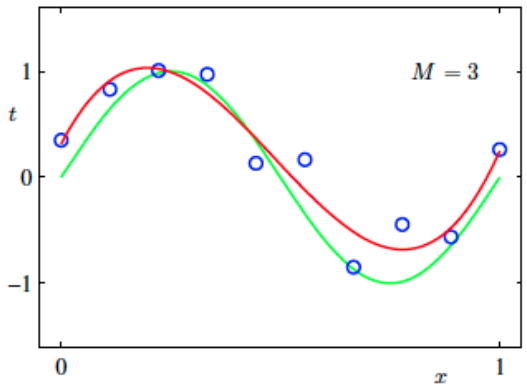
$$\mathcal{L}(x, t, \theta) = -\log_2 p(t|x, \theta)$$

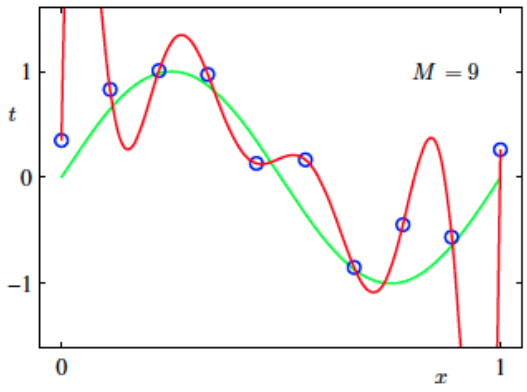
Проблема 1. Переобучение

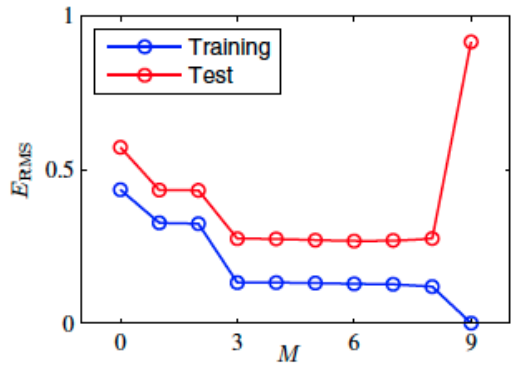
Задача

Аппроксимировать обучающую выборку полиномом M степени





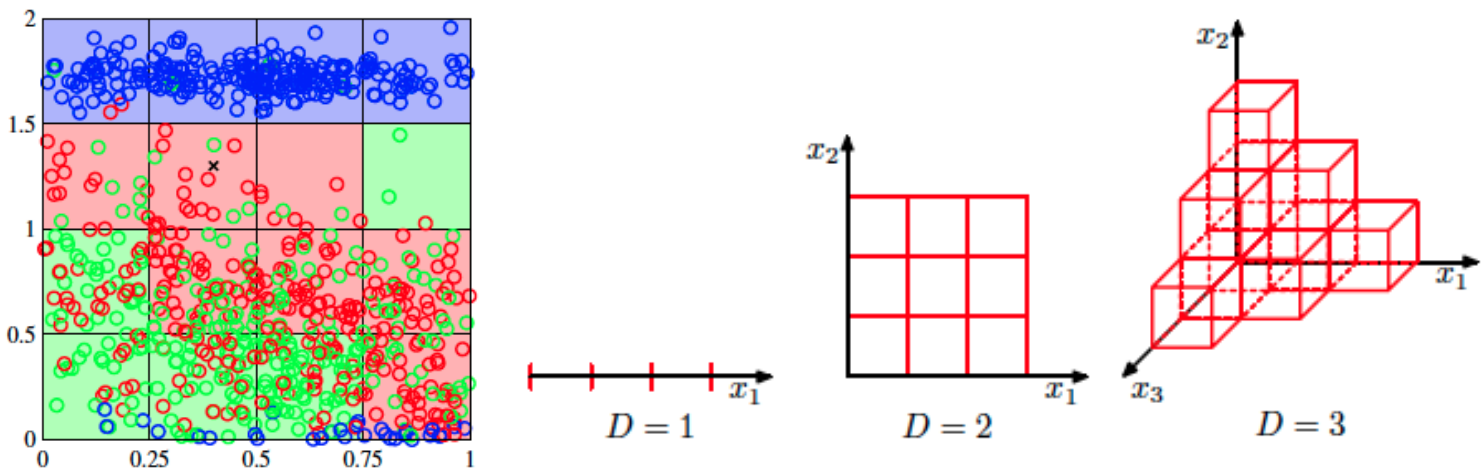




Проблема 2. Проклятие размерности

Задача

Классифицировать объекты.



Выбор модели

M Выдвигаем гипотезу насчет **модели** - семейства параметрических функций вида

$$Y = \{y(x, \theta) : X \times \Theta \to T\},$$

которая могла бы решить нашу задачу (model selection)

L Выбираем наилучшие параметры модели θ^*, используя **алгоритм обучения**

$$A(\mathbf{X}, \mathbf{T}) : (X, T)^N \to Y$$

(learning/inference)

D Используя полученную модель $y^*(x) = y(x, \theta^*)$, классифицируем неизвестные объекты (decision making)

Как оценить различные модели?

Идея

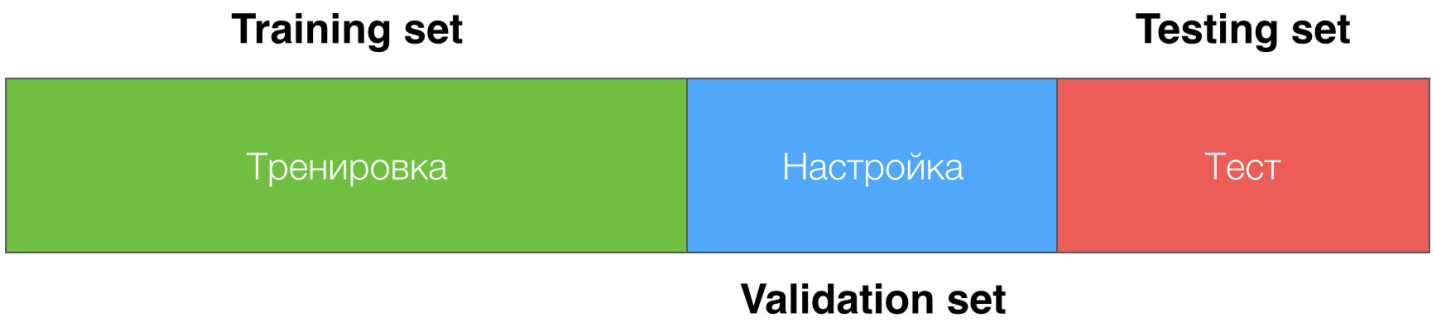
использовать долю неверно классифицированных объектов (error rate)

Важное замечание

error rate на обучающей выборке **НЕ** является хорошим показателем качества модели

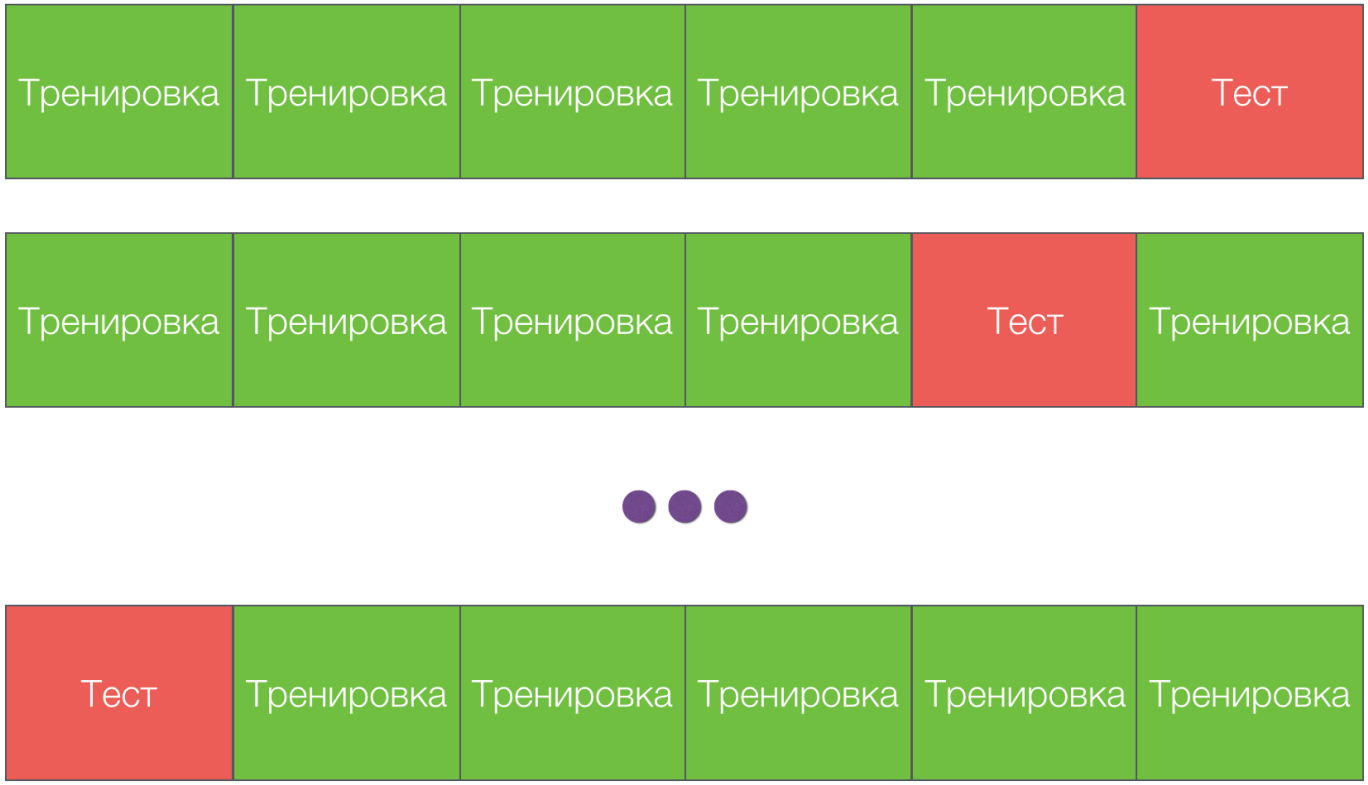
Решение 1: разделение выборки

Делим обучающую выборку на **тренировочную**, **валидационную** и **тестовую**



Решение 2: скользящий контроль

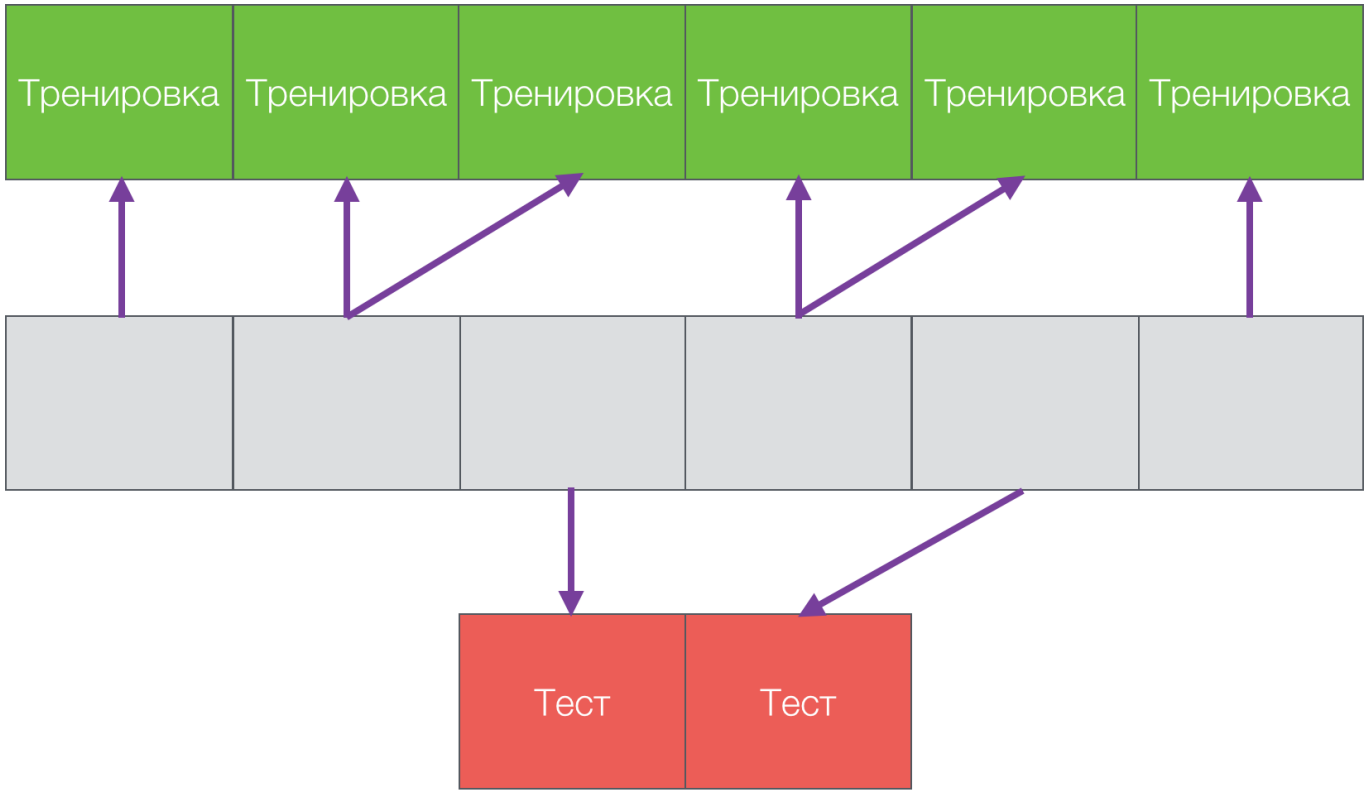
(n-times) (stratified) cross-validation



частный случай: leave-one-out

Решение 3: bootstrap

выбираем в тренировочную выбоку n объектов с возвращением



упражнение: найти математическое ожидание размера тестовой выборки.

Доверительный интервал для success rate

При тестировании на $N = 100$ объектах было получено 25 ошибок. Таким образом измеренная вероятность успеха (success rate) составила $f = 0.75$. Найти доверительный интервал для действительной вероятности успеха с уровнем доверия $\alpha = 0.8$.

Решение

Пусть p – действительная вероятность успеха в испытаниях бернулли, тогда

$$f \sim \mathcal{N}\left(p, p(1-p)/N\right).$$

Воспользовавшись табличным значением $P(-z \leq \mathcal{N}(0,1) \leq z) = \alpha$, имеем

$$P\left(-z \leq \frac{f-p}{\sqrt{p(1-p)/N}} \leq z\right) = \alpha,$$

откуда

$$p \in \left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) / \left(1 + \frac{z^2}{N}\right) = [0.69, 0.80]$$

Метрики качества. Вероятностные модели.

Пусть t_i - действительный класс для объекта x_i

- Information loss

$$-\frac{1}{N}\sum_i \log_2 p(t_i|x_i)$$

- Quadratic loss

$$\frac{1}{N}\sum_j (p(t_j|x_i) - a_j(x_i))^2,$$

где

$$a_j(x_i) = \begin{cases} 1, \text{ если } C_j = t_i \\ 0, \text{ иначе} \end{cases}$$

Метрики качества. Функции решения.

		Предсказанный	
		true	**false**
Действительный	**true**	TP	FN
	false	FP	TN

$$success\ rate = accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

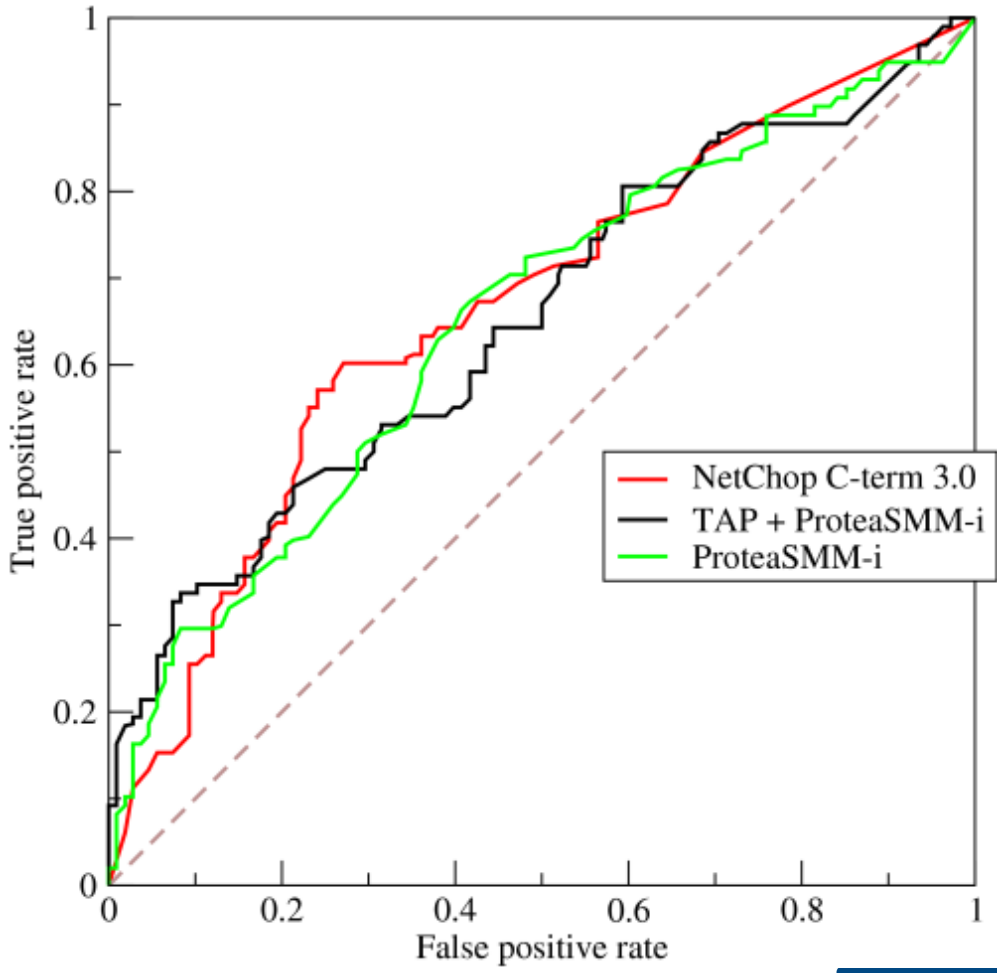
$$recall = TPR = \frac{TP}{TP + FN}; \quad precision = \frac{TP}{TP + FP}$$

$$FPR = \frac{FP}{FP + TN}$$

$$affinity = lift = \frac{accuracy}{p}$$

Receiver Operating Characteristic

$$TPR = \frac{TP}{TP + FN}; \quad FPR = \frac{FP}{FP + TN}$$



Упражнение

Простые классификаторы

В генеральной совокупности существуют объекты 3 классов, вероятность появления которых $p_1 < p_2 < p_3$. Первый классификатор относит все объекты к классу с большей вероятностью (то есть к третьему). Второй классификатор случайно относит объект к одному из классов в соответствии с базовым распределением. Рассчитать precision и recall, которые эти классификаторы дают для каждого из 3 классов.

Метрики качества. Регрессия

$$MSE = \frac{1}{N}\sum(y(x_i) - t_i)^2, \quad RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{N}\sum|y(x_i) - t_i|, \quad RMAE = \sqrt{MAE}$$

$$RSE = \frac{\sum(y(x_i) - t_i)^2}{\sum(t_i - \bar{t})^2}$$

$$correlation = \frac{S_{ty}}{S_t S_y}; \quad S_{ty} = \frac{\sum(y(i) - \overline{y(i)})(t_i - \bar{t})}{N-1}$$

$$S_y = \frac{\sum(y(i) - \overline{y(i)})^2}{N-1}; \quad S_t = \frac{\sum(t_i - \bar{t})^2}{N-1}$$

MDL принцип: интуиция

Спасибо!

Обратная связь