Лекционные слайды по курсу

Введение в

Машинное обучение

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml

Глава 14:

Оценочная и сравнительная классификация алгоритмов

Введение

- Вопросы:
  - Оценка ожидаемой ошибки обучающего алгоритма: будет ли погрешность метода ближайшего соседа (k-NN) при k=1 (1-NN) менее 2%?
  - Сравнение ожидаемых ошибок двух алгоритмов: метод ближайшего соседа (k-NN) более точен чем многослойный персептрон (MLP) ?
- Обучение/проверка/тестирование выборки
- Методы ресамплинга: Кросс-валидация по K блокам (*K*-fold cross-validation)

Предпочтение алгоритма

- Критерии (зависимые от применения):
  - Ошибка неверной классификации, или риск (функции потерь)
  - Время обучения/ уровень сложности
  - Время тестирования/ уровень сложности
  - Интерпретируемость
  - Лёгкая программируемость
- Обучение с учетом издержек классификации (Cost-sensitive learning)

Ресамплинг и кросс-валидация по K блокам

- Необходимость многократного обучения/ оценки выборки $\{X_i,V_i\}_i$: Обучения/оценки выборка из *i блоков*
- Кросс-валидация по K блокам: разделение X на *k блоков*, X_i, *i*=1,...,*K*

$$V_1 = X_1 \quad T_1 = X_2 \cup X_3 \cup \cdots \cup X_K$$
$$V_2 = X_2 \quad T_2 = X_1 \cup X_3 \cup \cdots \cup X_K$$
$$\vdots$$
$$V_K = X_K \quad T_K = X_1 \cup X_2 \cup \cdots \cup X_{K-1}$$

- T_i делится на *K*-2 частей

Кросс-валидация 5×2

- Кросс-валидация: 5 раз по 2 блока (Dietterich, 1998)

$$T_1 = X_1^{(1)} \quad V_1 = X_1^{(2)}$$
$$T_2 = X_1^{(2)} \quad V_2 = X_1^{(1)}$$

$$T_3 = X_2^{(1)} \quad V_3 = X_2^{(2)}$$
$$T_4 = X_2^{(2)} \quad V_4 = X_2^{(1)}$$

⋮

$$T_9 = X_5^{(1)} \quad V_9 = X_5^{(2)}$$
$$T_{10} = X_5^{(2)} \quad V_{10} = X_5^{(1)}$$

Выборка с возвратом (Bootstrapping)

- Возьмём образцы из выборки с последующим их возвратом
- Вероятность того, что мы не выберем образец после N попыток

$$\left(1-\frac{1}{N}\right)^{N} \approx e^{-1} = 0.368$$

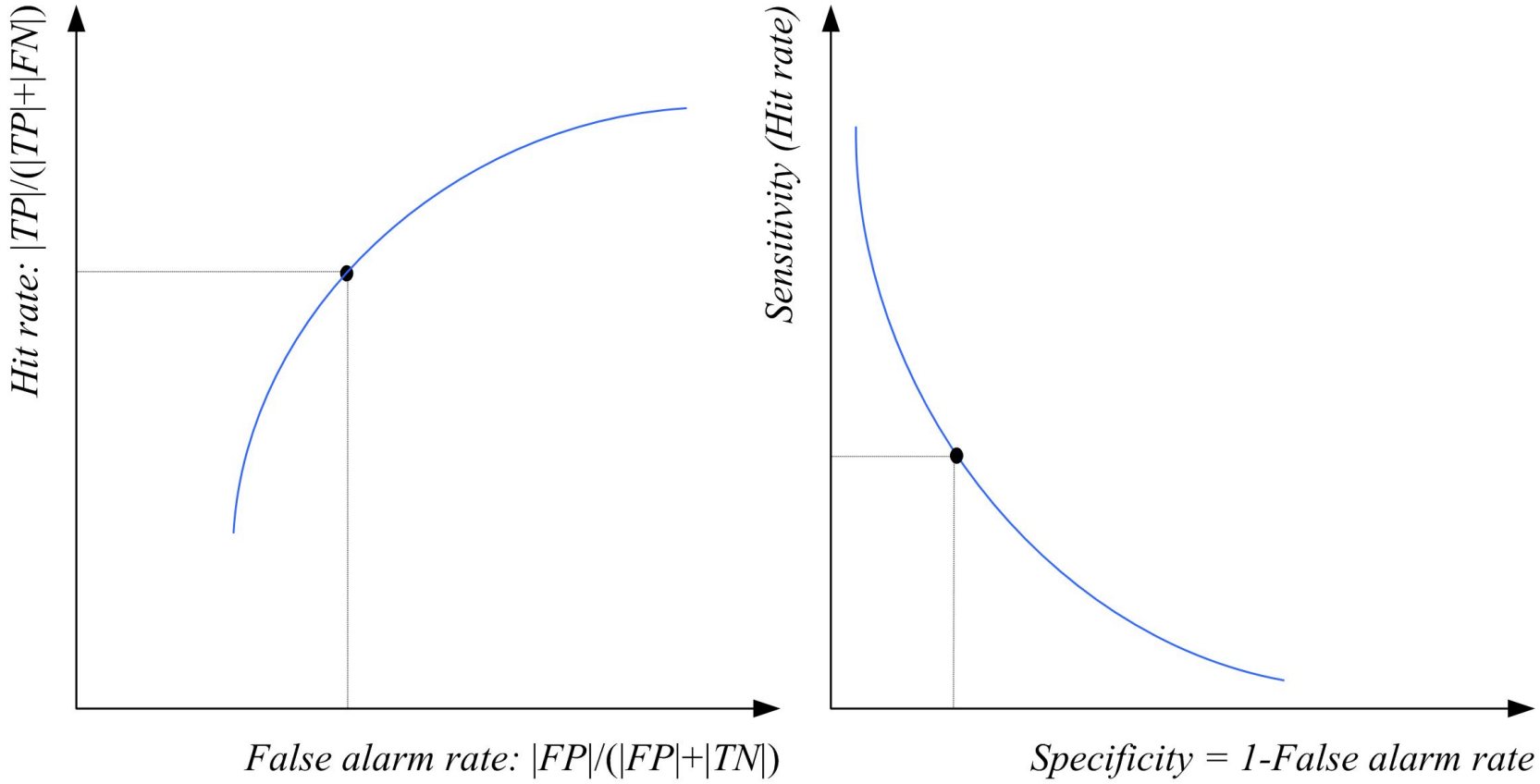
то есть, только 36.8% образцов являются новыми!

Ошибки измерения

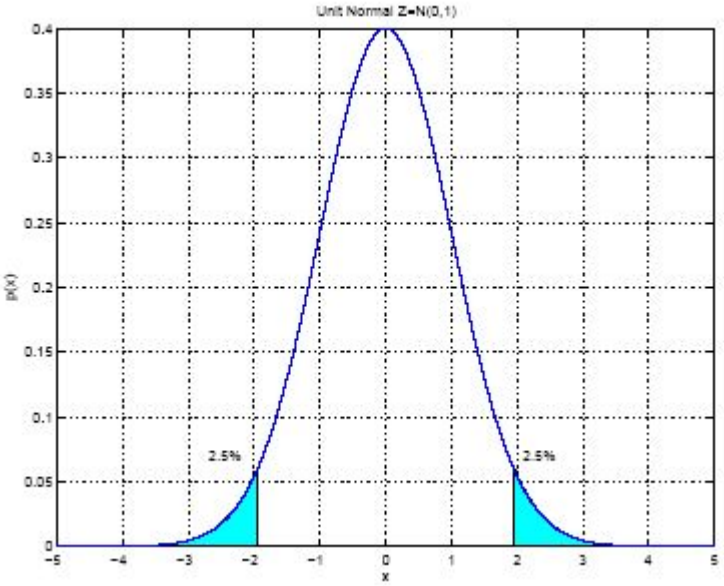
Исходный класс	Спрогнозированный класс	
	Yes	No
Yes	TP: True Positive	FN: False Negative
No	FP: False Positive	TN: True Negative

- Уровень ошибки = (FN+FP) / N
- Возврат = TP / (TP+FN) = чувствительность = коэффициент совпадений
- Точность = TP / (TP+FP)
- Определенность = TN / (TN+FP)
- Частота ложных тревог = FP / (FP+TN) = 1 - Specificity

Кривая ошибок (*ROC Curve*)



Интервальная оценка



- $X = \{ x^t \}_t$ where $x^t \sim N(\mu, \sigma^2)$
- $m \sim N(\mu, \sigma^2/N)$

$$\sqrt{N}\frac{(m-\mu)}{\sigma} \sim Z$$

$$P\left\{-1.96 < \sqrt{N}\frac{(m-\mu)}{\sigma} < 1.96\right\} = 0.95$$

$$P\left\{m - 1.96\frac{\sigma}{\sqrt{N}} < \mu < m + 1.96\frac{\sigma}{\sqrt{N}}\right\} = 0.95$$

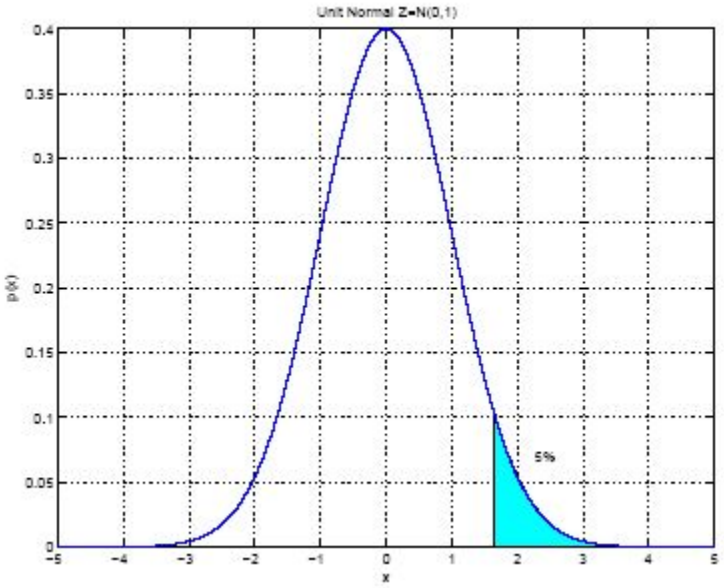
$$P\left\{m - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < m + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right\} = 1 - \alpha$$

100(1- α) % доверительный интервал

$$P\left\{\sqrt{N}\frac{(m-\mu)}{\sigma}<1.64\right\}=0.95$$

$$P\left\{m-1.64\frac{\sigma}{\sqrt{N}}<\mu\right\}=0.95$$

$$P\left\{m-z_{\alpha}\frac{\sigma}{\sqrt{N}}<\mu\right\}=1-\alpha$$



Когда σ^2 не известна:

$$S^2=\sum_t\left(x^t-m\right)^2/(N-1)$$

$$\frac{\sqrt{N}(m-\mu)}{S}\sim t_{N-1}$$

$$P\left\{m-t_{\alpha/2,N-1}\frac{S}{\sqrt{N}}<\mu<m+t_{\alpha/2,N-1}\frac{S}{\sqrt{N}}\right\}=1-\alpha$$

Проверка гипотезы

- Отклоняем недостоверную гипотезу если она не подкреплена выборкой с достаточной достоверностью
- $X = \{ x^t \}_t$ где $x^t \sim N (\mu, \sigma^2)$

  $H_0: \mu = \mu_0$ или $H_1: \mu \neq \mu_0$

  Принимаем H_0 с уровнем значимости α если μ_0 находится в $100(1- \alpha)$ доверительном интервале

$$\frac{\sqrt{N}(m-\mu_0)}{\sigma} \in \left(-z_{\alpha/2}, z_{\alpha/2}\right)$$

  Двусторонний тест

	Решение	
Правильность	Принять	Отклонить
True	Correct	Type I error
False	Type II error	Correct (Power)

- Односторонняя проверка: H_0: $\mu \leq \mu_0$ или H_1: $\mu > \mu_0$
  Принимаем если

$$\frac{\sqrt{N}(m-\mu_0)}{\sigma} \in (-\infty, z_\alpha)$$

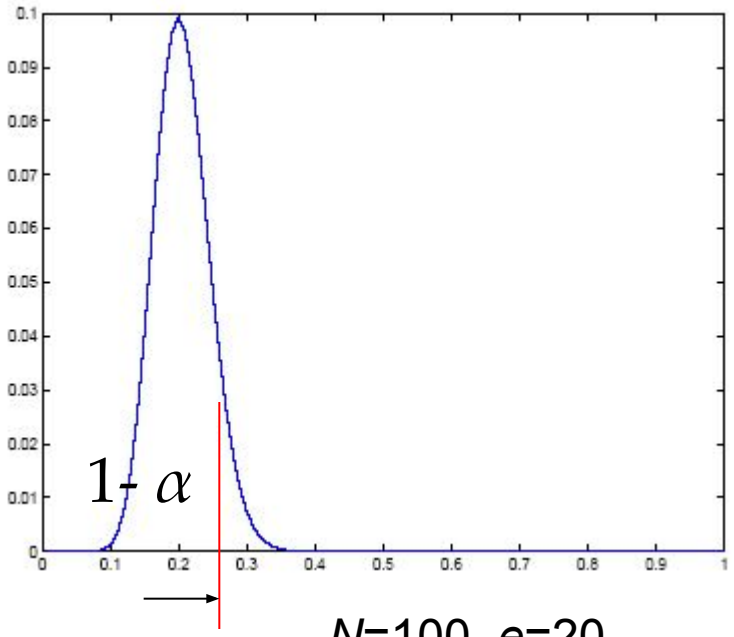
- Дисперсия неизвестна: используем t, вместо z
  Принимаем H_0: $\mu = \mu_0$ если

$$\frac{\sqrt{N}(m-\mu_0)}{S} \in (-t_{\alpha/2,N-1}, t_{\alpha/2,N-1})$$

Оценка ошибки: $H_0: p \leq p_0$ или $H_1: p > p_0$

- Одиночная обучающая/оценочная выборка: биноминальный тест

  Если вероятность ошибки p_0, то вероятность того что в N проверочных выборок будет e или менее ошибок



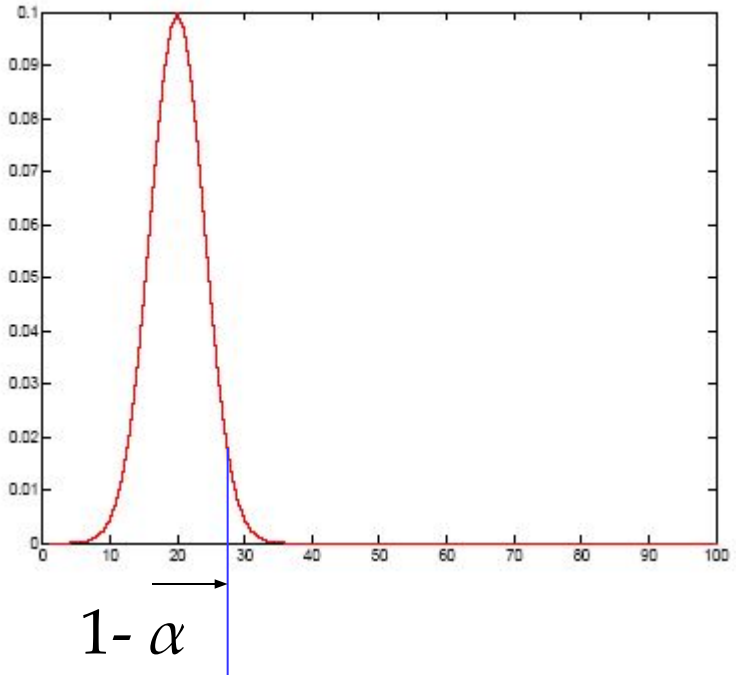
N=100, e=20

$$P\{X \leq e\} = \sum_{j=1}^{e} \binom{N}{j} p_0^{\,j} \left(1 - p_0^{\,j}\right)^{N-j}$$

Принимается если эта вероятность меньше чем1- α

Нормальная аппроксимация биноминального распределения

- Количество ошибок X приблизительно равно N со средним значением Np_0 и дисперсией $Np_0(1-p_0)$



$$\frac{X - Np_0}{\sqrt{Np_0(1-p_0)}} \sim \mathrm{Z}$$

Принимается, если вероятность того, что $X = e$ меньше чем $z_{1-\alpha}$

t-тест для парных выборок

- Многократное обучение/оценка выборки
- x_i^t = 1 если объект *t* неверно классифицирован на свертке *i*
- Уровень ошибки на свертке *i*: $p_i = \frac{\sum_{t=1}^{N} x_i^t}{N}$

- С помощью m и s^2 - математического ожидания и дисперсии p_i мы принимаем p_0 или меньшую ошибку если

$$\frac{\sqrt{K}(m - p_0)}{S} \sim t_{K-1}$$

меньше чем $t_{\alpha,K\text{-}1}$

Сравнительные классификаторы:

$H_0: \mu_0 = \mu_1$ или $H_1: \mu_0 \neq \mu_1$

- Однократное обучение/оценка выборки: тест Мак-Нимара

e_{00}: **неправильная классификация для обоих примеров**	e_{01}: **неправильная классификация для 1, но не для 2**
e_{10}: **неправильная классификация для 2, но не для 1**	e_{11}: **корректна для обоих примеров**

- Под H_0, мы понимаем $e_{01} = e_{10} = (e_{01} + e_{10})/2$

Принимаем если< $X^2_{\alpha,1}$

$$\frac{\left(|e_{01} - e_{10}| - 1\right)^2}{e_{01} + e_{10}} \sim \chi^2_1$$

Кросс-валидация по K блокам CV t-тест для парных выборок

- Используем кросс-валидацию по K блокам cv чтобы получить *K* обучающих/оценочных блоков
- p_i^1, p_i^2:Ошибки функции классификации 1 и 2 на блоке *i*
- $p_i = p_i^1 - p_i^2$: спаренное различие на блоке *i*
- Нулевая гипотеза это имеет ли p_i среднее значение,

$$H_0 : \mu = 0 \text{ или } H_0 : \mu \neq 0$$

$$m = \frac{\sum_{i=1}^{K} p_i}{K} \qquad s^2 = \frac{\sum_{i=1}^{K} (p_i - m)^2}{K-1}$$

$$\frac{\sqrt{K}(m-0)}{s} = \frac{\sqrt{K} \cdot m}{s} \sim t_{K-1}$$ принимаем, если находится в интервале $\left(-t_{\alpha/2,K-1}, t_{\alpha/2,K-1}\right)$

Кросс-валидация 5×2 cv t-тест для парных выборок

- Используем 5×2 cv чтобы получить 2 блока из 5 обучений/оценок репликаций (Dietterich, 1998)
- $p_i^{(j)}$: разница между ошибками 1 и 2 на блоках j=1, 2 репликации i=1,...,5

$$\bar{p}_i = \left(p_i^{(1)} + p_i^{(2)}\right)/2 \qquad s_i^2 = \left(p_i^{(1)} - \bar{p}_i\right)^2 + \left(p_i^{(2)} - \bar{p}_i\right)^2$$

$$\frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^{5} s_i^2 / 5}} \sim t_5$$

Двусторонняя проверка: принимаем H_0: $\mu_0 = \mu_1$ если находится в интервале $(-t_{\alpha/2,5}, t_{\alpha/2,5})$

Односторонняя проверка: принимаем H_0: $\mu_0 \leq \mu_1$ если $< t_{\alpha,5}$

Кросс-валидация 5×2 cv F-тест для парных выборок

$$\frac{\sum_{i=1}^{5}\sum_{j=1}^{2}\left(p_i^{(j)}\right)^2}{2\sum_{i=1}^{5}s_i^2} \sim F_{10,5}$$

Двусторонняя проверка: Принимаем H_0: $\mu_0 = \mu_1$ если < $F_{\alpha,10,5}$

Сравнение L>2 алгоритмов: анализ отклонений (Дисперсионный анализ)

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_L$$

- Ошибки *L* алгоритмов на *K* блоках

$$X_{ij} \sim \mathrm{N}\left(\mu_j, \sigma^2\right), \; j = 1, \ldots, L, \; i = 1, \ldots, K$$

- Мы строим две оценочные функции для σ^2 .

  Первая действительна если H_0 истина, другая действительна всегда.

  Мы отклоняем H_0 если две оценочные функции не совпадают.

Если H_0 верна:

$$m_j = \sum_{i=1}^{K} \frac{X_{ij}}{K} \sim \mathcal{N}\left(\mu, \sigma^2 / K\right)$$

$$m = \frac{\sum_{j=1}^{L} m_j}{L} \qquad S^2 = \frac{\sum_j \left(m_j - m\right)^2}{L-1}$$

Таким образом оценочной функцией для σ^2 является $K \cdot S^2$, а именно

$$\hat{\sigma}^2 = K \sum_{j=1}^{L} \frac{\left(m_j - m\right)^2}{L-1}$$

$$\sum_j \frac{\left(m_j - m\right)^2}{\sigma^2 / K} \sim \chi^2_{L-1} \qquad SSb \equiv K \sum_j \left(m_j - m\right)^2$$

И когда H_0 верно, мы получим

$$\frac{SSb}{\sigma^2} \sim \chi^2_{L-1}$$

В не зависимости от H_0 наша вторая оценочная функция для σ^2 это среднее значение групповой дисперсии S_j^2 :

$$S_j^2 = \frac{\sum_{i=1}^{K} \left(X_{ij} - m_j\right)^2}{K-1} \qquad \hat{\sigma}^2 = \sum_{j=1}^{L} \frac{S_j^2}{L} = \sum_j \sum_i \frac{\left(X_{ij} - m_j\right)^2}{L\left(K-1\right)}$$

$$SSw \equiv \sum_j \sum_i \left(X_{ij} - m_j\right)^2$$

$$\left(K-1\right)\frac{S_j^2}{\sigma^2} \sim \chi_{K-1}^2 \qquad \frac{SSw}{\sigma^2} \sim \chi_{L(K-1)}^2$$

$$\left(\frac{SSb/\sigma^2}{L-1}\right) \Big/ \left(\frac{SSw/\sigma^2}{L\left(K-1\right)}\right) = \frac{SSb/\left(L-1\right)}{SSw/\left(L\left(K-1\right)\right)} \sim F_{L-1,L(K-1)}$$

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_L \text{ if } < F_{\alpha,L-1,L(K-1)}$$

Другие тесты

- Оценка диапазона (Ньюмена-Койлса) $\frac{1}{1}\frac{45}{23}$
- Непараметрические тесты (Критерий знаков, Крускала — Уоллиса)
- Противопоставление: Проверить отличаются ли 1 и 2 от 3,4, и 5
- Множественные сравнения требуют Поправки Бонферрони если для достижения уровня α есть m гипотез, то каждая из них должна иметь значение α/m.
- Регрессия: Центральная предельная теорема утверждает что сумма независимых одинаково распределённых случайных величин переменных любого распределения есть величина приблизительно нормальная и к ней могут быть применены предшествующие методы