Машинное обучение

Лекция 11. Линейная регрессия. PCA.

Катя Тузова

Разбор летучки

Метод наименьших квадратов

В первом домашнем задании мы реализовывали метод наименьших квадратов.
К какому типу классификаторов он относится?

Регрессия

X– объекты в $\mathbb{R}^n$; Y — ответы в $\mathbb{R}$
$X^l = (x_i, y_i)_{i=1}^l$ – обучающая выборка
$y_i = y(x_i)$, $y : X \to Y$ – неизвестная зависимость

$a(x) = f(x, w)$ – модель зависимости,
$w \in \mathbb{R}^p$ – вектор параметров модели.

Метод наименьших квадратов (МНК):

$$Q(w, X^l) = \sum_{i=1}^{l} \alpha_i (f(x_i, w) - y_i)^2 \to \min_w$$

где α_i – вес, степень важности i-го объекта.
$Q(w^*, X^l)$ — остаточная сумма квадратов

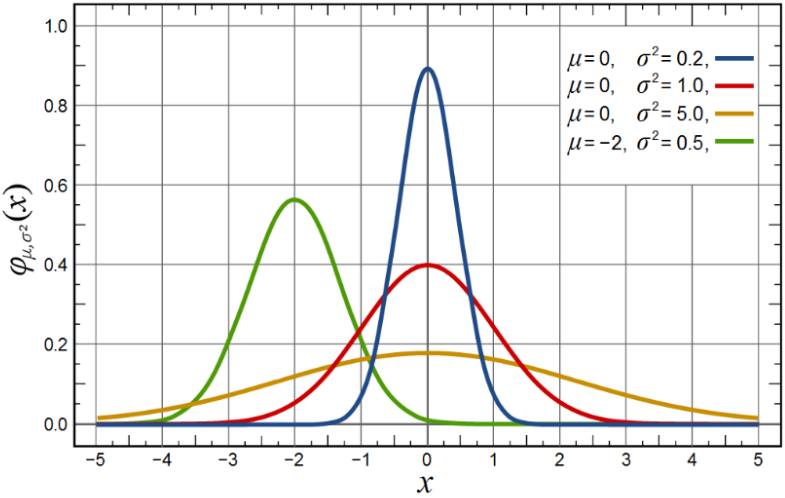
Метод максимума правдоподобия

Модель данных с некоррелированным гауссовским шумом:

$$y(x_i) = f(x_i, w) + \varepsilon_i, \qquad \varepsilon_i \in N(0, \sigma_i^2), i = 1, \dots, l$$

Вопрос: Как выглядит плотность одномерного Гауссовского распределения?

Нормальное распределение



Метод максимума правдоподобия

Метод максимума правдоподобия (ММП):

$$L(\varepsilon_1, \ldots, \varepsilon_l | w) = \prod_{i=1}^{l} p(\varepsilon_i) = \prod_{i=1}^{l} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma_i^2}\right) \to \max_w$$

$$-\ln L(\varepsilon_1, \ldots, \varepsilon_l | w) = const(w) + \frac{1}{2} \sum_{i=1}^{l} \frac{(f(x_i, w) - y_i)^2}{\sigma_i^2} \to \min_w$$

Метод максимума правдоподобия

Метод максимума правдоподобия (ММП):

$$-\ln L(\varepsilon_1, \dots, \varepsilon_l | w) = const(w) + \frac{1}{2} \sum_{i=1}^{l} \frac{(f(x_i, w) - y_i)^2}{\sigma_i^2} \to \min_w$$

Метод наименьших квадратов:

$$Q(w, X^l) = \sum_{i=1}^{l} \alpha_i (f(x_i, w) - y_i)^2 \to \min_w$$

Удивительный факт: Постановки МНК и ММП, совпадают, причём веса объектов обратно пропорциональны дисперсии шума, $\alpha_i = \sigma_i^{-2}$

Многомерная линейная регрессия

$f_1(x), \dots, f_n(x)$ – числовые признаки
Модель многомерной линейной регрессии:

$$f(x, w) = \sum_{j=1}^{n} w_j f_j(x), \qquad w \in \mathbb{R}$$

Функционал квадрата ошибки:

$$Q(w, X^l) = \sum_{i=1}^{l} (f(x_i, w) - y_i)^2 \to \min_w$$

Матричное представление

$$F_{l\times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix} \; y_{l\times 1} = \begin{pmatrix} y_1 \\ \dots \\ y_l \end{pmatrix} \; w_{n\times 1} = \begin{pmatrix} w_1 \\ \dots \\ w_n \end{pmatrix}$$

Функционал квадрата ошибки:

$$Q(w, X^l) = \sum_{i=1}^{l} (f(x_i, w) - y_i)^2 = \|Fw - y\|^2 \to \min_w$$

Нормальная система уравнений

Необходимое условие минимума:

$$\frac{\partial Q(w)}{\partial w} = 2F^T(Fw - y) = 0$$

Откуда следует нормальная система задачи МНК:

$$F^T F w = F^T y$$

$F^T F$ – ковариационная матрица признаков $f_1, \dots, f_n$

Нормальная система уравнений

Нормальная система задачи МНК:

$$F^T F w = F^T y$$

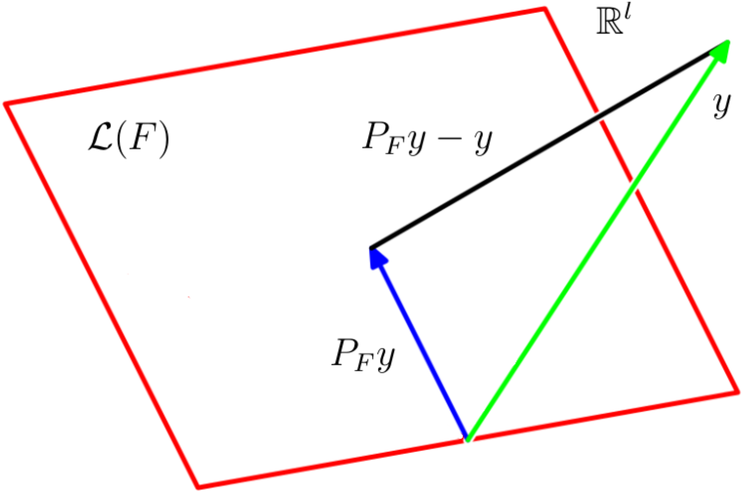
Решение системы:

$$w^* = (F^T F)^{-1} F^T y = F^+ y$$

F^+ – псевдообратная матрица

Значение функционала: $Q(w^*) = \|P_F y - y\|^2$,
где $P_F = F F^+ = F(F^T F)^{-1} F^T$ – проекционная матрица

Геометрический смысл



Сингулярное разложение

Произвольная $l \times n$-матрица представима в виде сингулярного разложения:

$$F = VDU^T$$

Основные свойства сингулярного разложения:

- $V_{l\times n} = (v_1, \dots, v_n)$ ортогональна, $V^TV = I_n$,
  столбцы v_j - собственные векторы матрицы FF^T
- $U_{n\times n} = (u_1, \dots, u_n)$ ортогональна, $U^TU = I_n$,
  столбцы u_j - собственные векторы матрицы F^TF
- D диагональна, $D_{n\times n} = \mathrm{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$,
  $\lambda_j > 0$ - собственные значения матриц F^TF и FF^T

Решение МНК через сингулярное разложение

Псевдообратная F^+, вектор МНК-решения w^*, МНК-аппроксимация целевого вектора Fw^*

$$F^+ = (UDV^TVDU^T)^{-1}UDV^T = UD^{-1}V^T = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^T$$

$$w^* = F^+y = UD^{-1}V^Ty = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y)$$

$$Fw^* = P_F y = (VDU^T)UD^{-1}V^Ty = VV^Ty = \sum_{j=1}^{n} v_j (v_j^T y)$$

$$\|w^*\|^2 = \|D^{-1}V^Ty\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^T y)^2$$

Проблема мультиколлинеарности

Если $\exists \gamma \in \mathbb{R}^n : F\gamma \approx 0$, то некоторые λ_j близки к нулю.
Число обусловленности $n \times n$-матрицы $F^T F = A$:

$$\mu(A) = \|A\|\|A^{-1}\| = \frac{\max\limits_{u:\|u\|=1} \|Au\|}{\min\limits_{u:\|u\|=1} \|Au\|} = \frac{\lambda_{max}}{\lambda_{min}}$$

При умножении обратной матрицы на вектор, $z = A^{-1}u$, относительная погрешность усиливается в $\mu(A)$ раз:

$$\frac{\|\delta z\|}{\|z\|} \leq \mu(A) \frac{\|\delta u\|}{\|u\|}$$

Проблема мультиколлинеарности

Если матрица $F^T F$ плохо обусловлена, то:

- решение становится неустойчивым и неинтерпретируемым, $\|w^*\|$ велико
- на обучении $Q(w^*, X^l) = \|Fw^* - y\|$ – мало
- на контроле $Q(w^*, X^k) = \|F'w^* - y'\|$ – велико

Вопрос: Как бороться с этой проблемой?

Проблема мультиколлинеарности

Стратегии устранения мультиколлинеарности и переобучения:

- Отбор признаков: $f_1, \dots, f_n \to f_{j_1}, \dots, f_{j_m}, \quad m << n$
- Преобразование признаков: $f_1, \dots, f_n \to g_1, \dots, g_m, \quad m << n$
- Регуляризация: $\|w\| \to \min$

Гребневая регрессия

Штраф за увеличение нормы вектора весов $\|w\|$

$$Q_\tau(w) = \|Fw - y\|^2 + \frac{1}{\sigma}\|w\|^2$$

где $\tau = \frac{1}{\sigma}$ – неотрицательный параметр регуляризации.

Модифицированное МНК-решение (τI_n — «гребень»)

$$w_\tau^* = (F^T F + \tau I_n)^{-1} F^T y$$

Вопрос: Можно ли подбирать τ не вычисляя каждый раз обратную матрицу?

Преимущество сингулярного разложения

Модифицированное МНК-решение (τI_n — «гребень»)

$$w_\tau^* = (F^T F + \tau I_n)^{-1} F^T y$$

Преимущество сингулярного разложения:
можно подбирать параметр τ , вычислив сингулярное разложение только один раз.

Сингулярное разложение

$$w_\tau^* = U(D^2 + \tau I_n)^{-1} D V^T y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^T y)$$

$$F w_\tau^* = V D U^T w_\tau^* = V \operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^T y = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^T y)$$

$$\|w_\tau^*\|^2 = \|U(D^2 + \tau I_n)^{-1} D V^T y\|^2 = \sum_{j=1}^{n} \frac{\lambda_j}{(\lambda_j + \tau)^2} (v_j^T y)^2$$

$F w_\tau^* \neq F w^*$ – зато решение становится более устойчивым

Выбор параметра регуляризации τ

Контрольная выборка: $X^k = (x'_i, y'_i)_{i=1}^k$

$$Q(w^*_\tau, X^k) = \|F' w^*_\tau - y'\|^2 = \|F' U \operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^T y - y'\|^2$$

Зависимость $Q(\tau)$ обычно имеет характерный минимум.

Сокращение «эффективной размерности»

Сокращение весов:

$$\|w_\tau^*\|^2 = \sum_{j=1}^{n} \frac{\lambda_j}{(\lambda_j + \tau)^2} (v_j^T y)^2 < \|w^*\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^T y)^2$$

Роль размерности играет след проекционной матрицы:

$$trF(F^T F)^{-1} F^T = tr(F^T F)^{-1} F^T F = trI_n = n$$

При использовании регуляризации:

$$trF(F^T F + \tau I_n)^{-1} F^T = tr\ \operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} < n$$

LASSO

LASSO – Least Absolute Shrinkage and Selection Operator

$$\begin{cases} Q(w, X^l) = \|Fw - y\|^2 \to \min\limits_w \\ \sum_{j=1}^n |w_j| \le \kappa \end{cases}$$

После замены переменных:

$$\begin{cases} w_j = w_j^+ - w_j^- \\ |w_j| = w_j^+ + w_j^-, \quad w_j^+, w_j^- \ge 0 \end{cases}$$

ограничения принимают канонический вид:

$$\sum_{j=1}^n w_j^+ + w_j^- \le \kappa$$

Чем меньше κ, тем больше j таких, что $w_j^+ = w_j^- = 0$

Негладкие регуляризаторы

Elastic Net:

$$\frac{1}{2}\|Fw - y\|^2 + \mu \sum_{j=1}^{n} |w_j| + \frac{\tau}{2} \sum_{j=1}^{n} w_j^2 \to \min_w$$

Support Features Machine:

$$\frac{1}{2}\|Fw - y\|^2 + \tau \sum_{j=1}^{n} R_\mu(w_j) \to \min_w$$

$$R_\mu(w_j) = \begin{cases} 2\mu|w_j|, & |w_j| \le \mu \\ \mu^2 + w_j^2, & |w_j| \ge \mu \end{cases}$$

Метод главных компонент

$f_1(x), \dots, f_n(x)$ – исходные числовые признаки
$g_1(x), \dots, g_m(x)$ – новые числовые признаки, $m \times n$

Вопрос: Как сформулировать требование к новым признакам?

Метод главных компонент

$f_1(x), \dots, f_n(x)$ – исходные числовые признаки
$g_1(x), \dots, g_m(x)$ – новые числовые признаки, $m \times n$
Требование: старые признаки должны линейно восстанавливаться по новым:

$$\hat{f}_j(x) = \sum_{s=1}^{m} g_s(x) u_{js}, \quad j = 1, \dots, n, \quad \forall x \in X$$

как можно точнее на обучающей выборке $x_1, \dots, x_l$:

$$\sum_{i=1}^{l} \sum_{j=1}^{n} (\hat{f}_j(x_i) - f_j(x_i))^2 \to \min_{g_s(x_i), u_{js}}$$

Матричные обозначения

$$F_{l\times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix} \quad G_{l\times m} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_l) & \dots & g_m(x_l) \end{pmatrix}$$

$$U_{n\times m} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}$$

U – линейное преобразование новых признаков в старые

$$\hat{F} = GU^T \approx F$$

Найти: новые признаки G и преобразование U:

$$\sum_{i=1}^{l}\sum_{j=1}^{n}(\hat{f}_j(x_i) - f_j(x_i))^2 = \|GU^T - F\|^2 \to \min_{G,U}$$

Основная теорема

Если $m < \operatorname{rank} F$, то минимум $\|GU^T - F\|^2$ достигается, когда столбцы U - это с.в. матрицы $F^T F$, соответствующие m максимальным с.з. $\lambda_1, \dots, \lambda_m$, а матрица $G = FU$.

При этом:

- матрица U ортонормирована: $U^T U = I_m$
- матрица G ортогональна: $G^T G = \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_m)$
- $U\Lambda = F^T F U$, $G\Lambda = F F^T G$
- $\|GU^T - F\|^2 = \|F\|^2 - tr\Lambda = \sum_{j=m+1}^{n} \lambda_j$

Связь с сингулярным разложением

Если взять $m = n$, то:

- $\|GU^T - F\|^2 = 0$
- представление $\hat{F} = GU^T = F$ точное и совпадает с сингулярным разложением при $G = V\sqrt{\Lambda}$

$$F = GU^T = V\sqrt{\Lambda}U^T, \;\; U^T U = I_m, \;\; V^T V = I_m$$

- линейное преобразование U работает в обе стороны:

$$F = GU^T, \quad G = FU$$

Преобразование U называется декоррелирующим

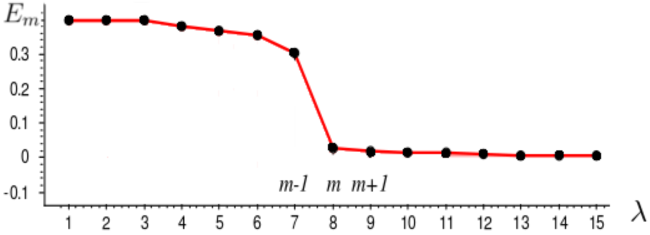
Эффективная размерность выборки

Упорядочим с.з. $F^T F$ по убыванию: $\lambda_1 > \cdots > \lambda_n > 0$

Эффективная размерность выборки – это наименьшее целое m, при котором

$$E_m = \frac{\|GU^T - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \cdots + \lambda_n}{\lambda_1 + \cdots + \lambda_n} \leq \varepsilon$$

Критерий «крутого склона»: находим $m : E_{m-1} >> E_m$:



Решение задачи НК в новых признаках

Заменим F на её приближение GU^T:

$$\|GU^T w - y\|^2 = \|G\hat{w} - y\|^2 \to \min_{\hat{w}}$$

Связь нового и старого вектора коэффициентов:

$$w = U\hat{w}, \qquad \hat{w} = U^T w$$

Решение задачи наименьших квадратов относительно $\hat{w}$ (единственное отличие – m слагаемых вместо n):

$$\hat{w}^* = D^{-1}V^T y = \sum_{j=1}^{m} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y)$$

$$G\hat{w}^* = VV^T y = \sum_{j=1}^{m} v_j (v_j^T y)$$

Нелинейная регрессия

Вопрос: Что изменится, если модель регрессии не линейна?

$$f(x, w), \qquad w \in \mathbb{R}^p$$

Метод Ньютона-Рафсена

Начальное приближение $w^{(0)} = (w_1^{(0)}, \ldots, w_p^{(0)})$
Итерационный процесс: $w^{(t+1)} = w^{(t)} - h_t(Q''(w^{(t)}))^{-1}Q'(w^{(t)})$

$Q'(w^{(t)})$ – градиент функционала Q в точке $w^{(t)}$
$Q''(w^{(t)})$ – гессиан функционала Q в точке $w^{(t)}$
h_t – величина шага

Метод Ньютона-Рафсена

$$Q(w, X^l) = \sum_{i=1}^{l} (f(x_i, w) - y_i)^2 \to \min_w$$

$$\frac{\partial Q(w)}{\partial w_j} = 2\sum_{i=1}^{l} (f(x_i, w) - y_i) \frac{\partial (f(x_i, w))}{\partial w_j}$$

$$\frac{\partial^2 Q(w)}{\partial w_j \partial w_k} = 2\sum_{i=1}^{l} \frac{\partial (f(x_i, w))}{\partial w_j} \frac{\partial (f(x_i, w))}{\partial w_k} - 2\sum_{i=1}^{l} (f(x_i, w) - y_i) \frac{\partial^2 (f(x_i, w))}{\partial w_j \partial w_k}$$

Вопрос: Какая часть самая тяжелая?

Метод Ньютона-Рафсена

$$\frac{\partial Q(w)}{\partial w_j} = 2\sum_{i=1}^{l}(f(x_i,w) - y_i)\frac{\partial(f(x_i,w))}{\partial w_j}$$

$$\frac{\partial^2 Q(w)}{\partial w_j \partial w_k} = 2\sum_{i=1}^{l}\frac{\partial(f(x_i,w))}{\partial w_j}\frac{\partial(f(x_i,w))}{\partial w_k} - 2\sum_{i=1}^{l}(f(x_i,w) - y_i)\frac{\partial^2(f(x_i,w))}{\partial w_j \partial w_k}$$

Линеаризация $f(x_i,w)$ в окрестности текущего $w^{(t)}$:

$$f(x_i,w) = f(x_i,w^{(t)}) + \sum_{j=1}^{p}\frac{\partial(f(x_i,w))}{\partial w_j}(w_j - w_j^{(t)}) + o(w_j - w_j^{(t)})$$

$\Rightarrow$ второе слагаемое в гессиане обнулилось

Матричные обозначения

$$F_t = \left(\frac{\partial(f(x_i,w))}{\partial w_j^{(t)}}\right)_{l\times p}$$ – матрица первых производных

$$f_t = \big(f(x_i, w^{(t)})\big)_{l\times 1}$$ – вектор значений f

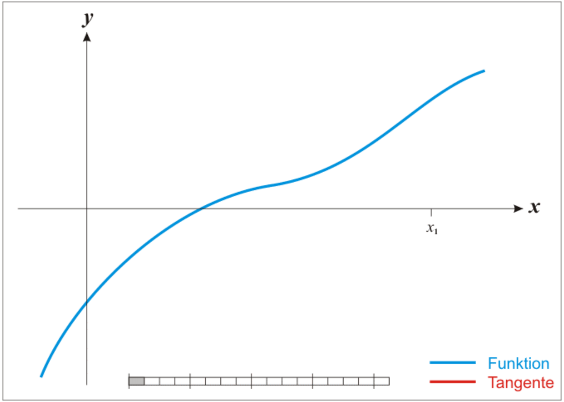
Формула t-й итерации метода Ньютона–Гаусса:

$$w^{(t+1)} = w^{(t)} - h_t \underbrace{\left(F_t^T F_t\right)^{-1} F_t^T (f_t - y)}_{\beta}$$

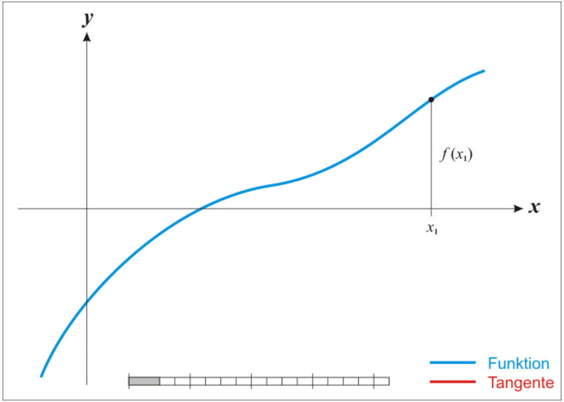
где β – решение многомерной линейной регрессии

$$\|F_t \beta - (f_t - y)\|^2 \to \min_{\beta}$$

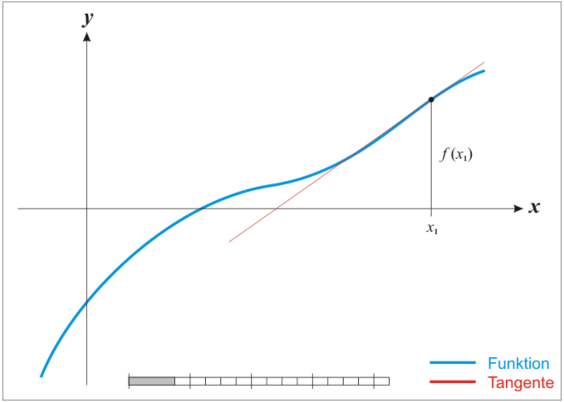
Метод Ньютона-Рафсена



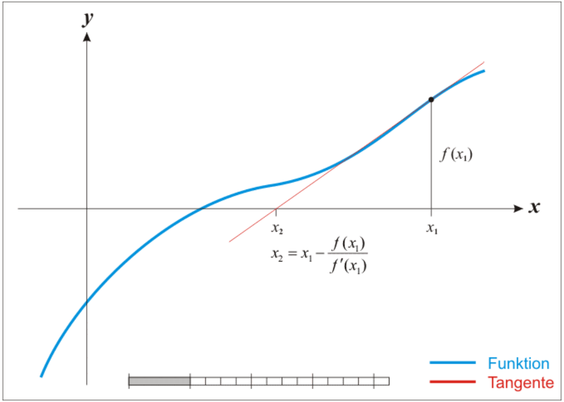
Метод Ньютона-Рафсена



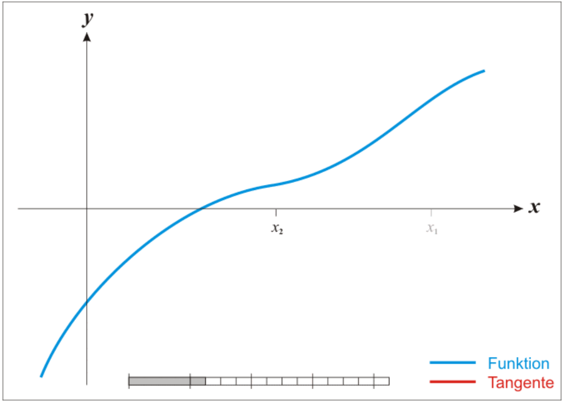
Метод Ньютона-Рафсена



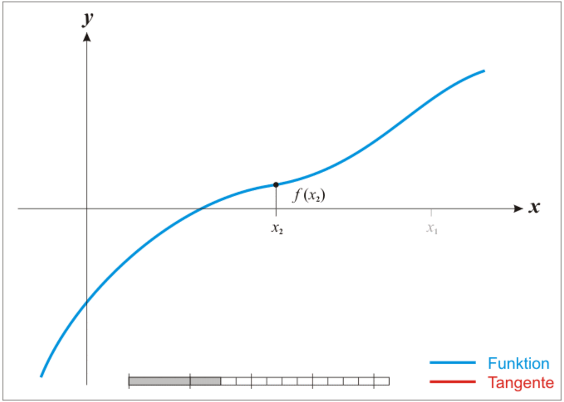
Метод Ньютона-Рафсена



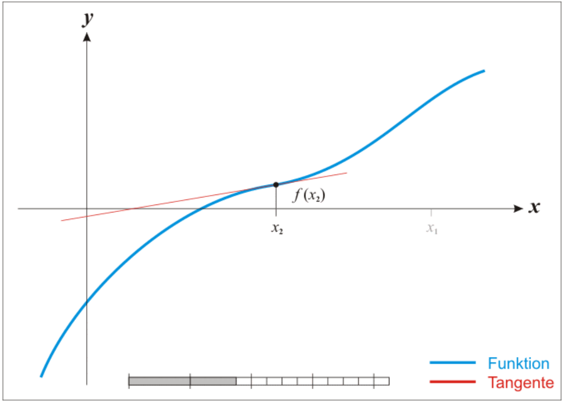
Метод Ньютона-Рафсена



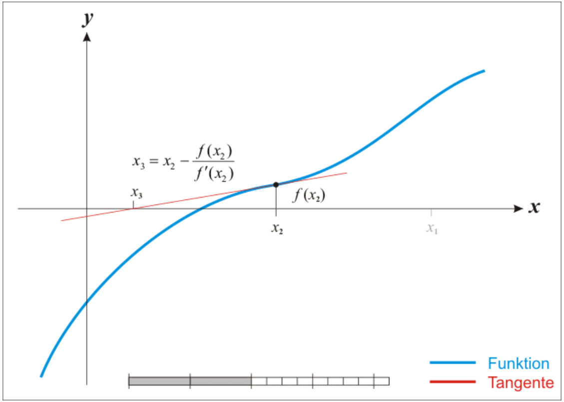
Метод Ньютона-Рафсена



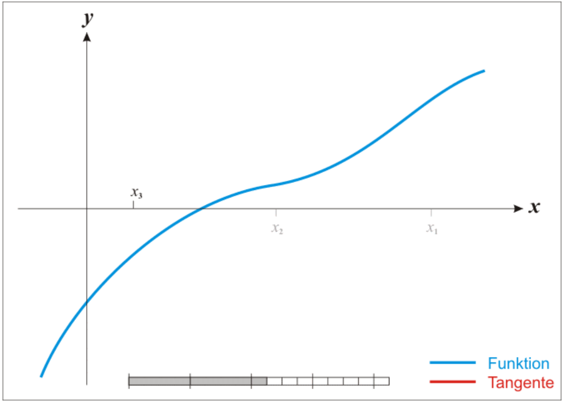
Метод Ньютона-Рафсена



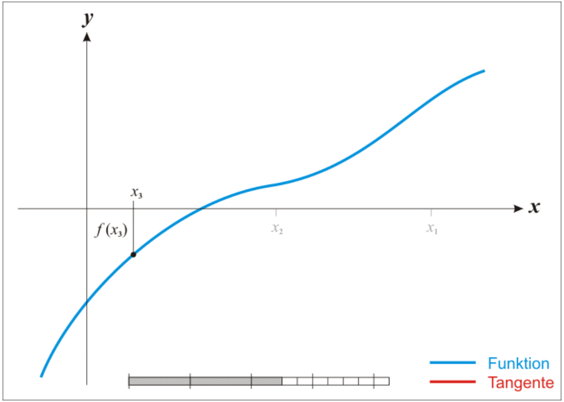
Метод Ньютона-Рафсена



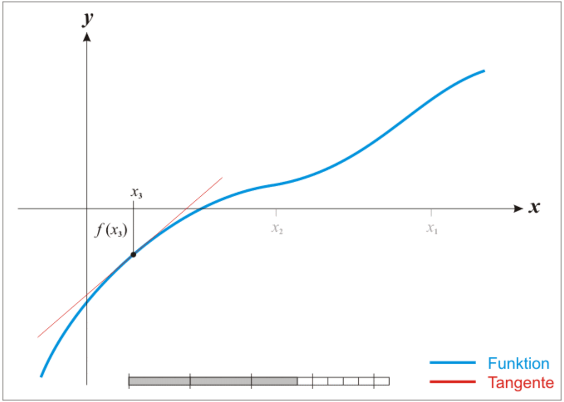
Метод Ньютона-Рафсена



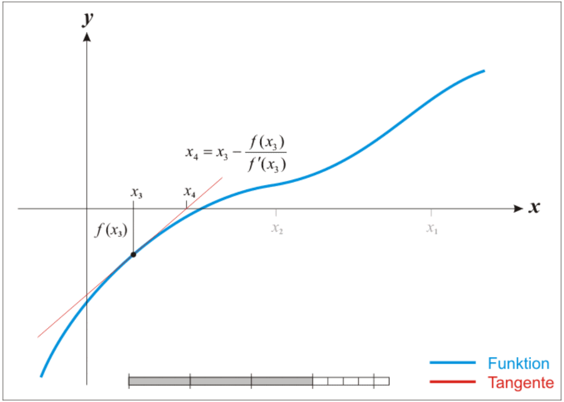
Метод Ньютона-Рафсена



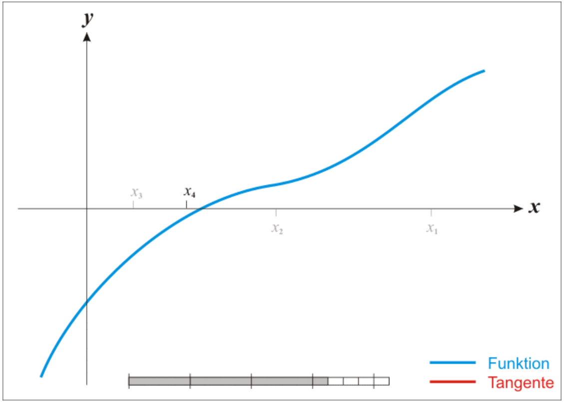
Метод Ньютона-Рафсена



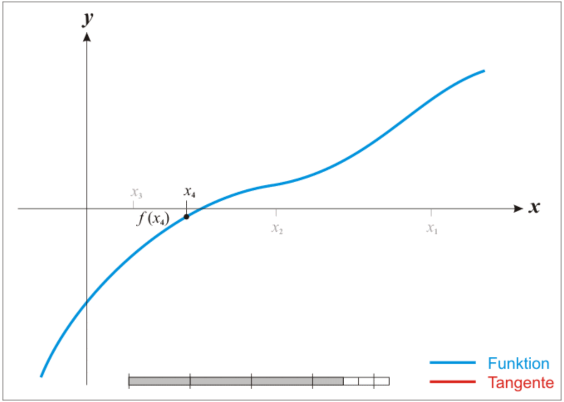
Метод Ньютона-Рафсена



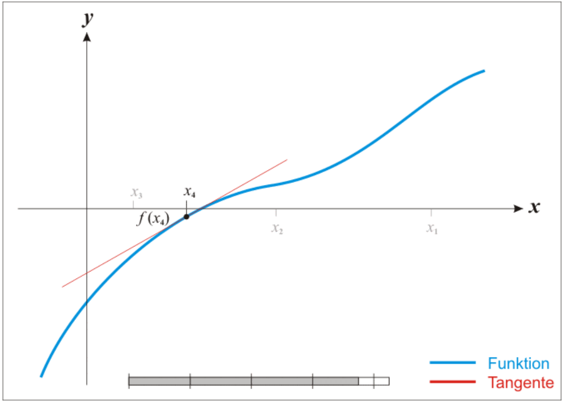
Метод Ньютона-Рафсена



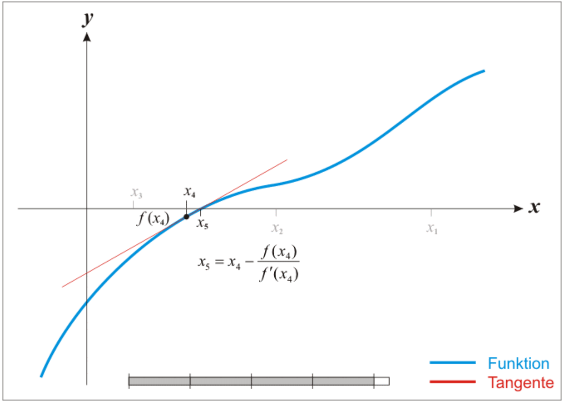
Метод Ньютона-Рафсена



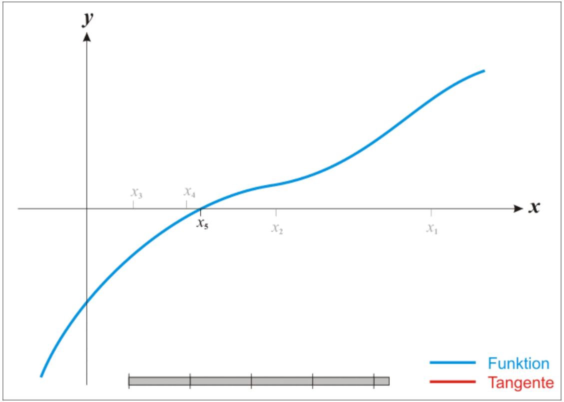
Метод Ньютона-Рафсена



Метод Ньютона-Рафсена



Метод Ньютона-Рафсена



На следующей лекции

- Нейронные сети