

Linear models: SVM, PCA

Vladislav Goncharenko

ML Teamlead, DZEN



MSU, spring 2024

Questions



1. Что предсказывает логистическая регрессия?
2. Формула сигмоиды
3. Назовите стратегии агрегации бинарных предсказаний в мультиклассовое. Какое количество моделей необходимо натренировать для их использования?
4. Что должна предсказывать модель чтобы мы могли подсчитать roc-auc?
5. Значение (семантика) метрики классификации
 - a. precision
 - b. recall

Recap

Lecture 3: Logistic Regression

- Linear classification
 - margin
 - loss functions
- Logistic regression
 - sigmoid derivation
 - Maximum Likelihood Estimation (MLE)
 - logistic loss
 - probability calibration
- Multiclass aggregation strategies
 - One vs Rest
 - One vs One
- Metrics in classification
 - Accuracy, Balanced accuracy
 - Precision, Recall, F-score
 - ROC curve, PR curve, AUC
 - Confusion matrix

Outline

- Support Vector Machine (SVM)
 - History
 - Motivation
 - Solution for separable design
 - Inseparable design, soft margin
 - Kernels
- Dimensionality reduction and PCA
 - Problem statement
 - Eckart–Young theorem
 - Equivalent definitions

Maximum Likelihood Estimation

girafe
ai

00

Maximum Likelihood Estimation



What are reasons behind defining “best” linear estimator?

Maximize probability of particular parameter to explain given data

$$L(\theta|X, Y) = P(X, Y|\theta)$$

assuming i.i.d. observations

$$P(X, Y|\theta) = \prod_{i=1}^n P(x^i, y^i|\theta)$$

$$\log L(\theta|X, Y) = \sum_{i=1}^n \log P(x^i, y^i|\theta)$$

Support Vector Machine

girafe
ai

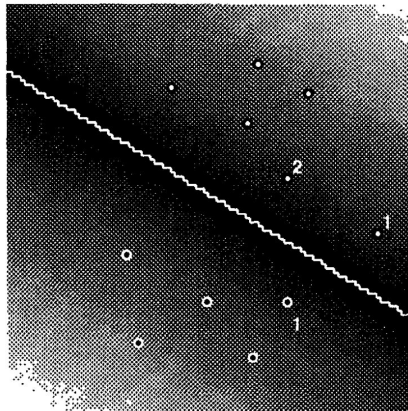
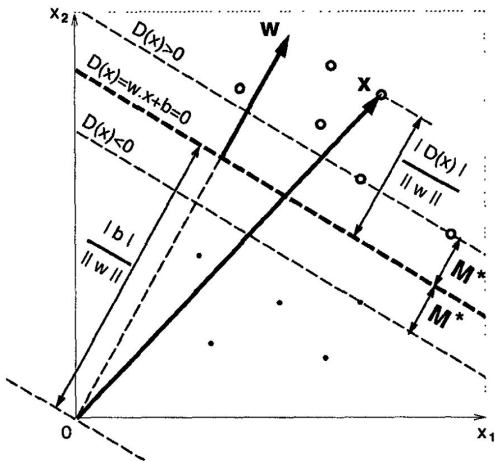
01

Support Vector Machine



1. History
2. Motivation
3. Solution for separable design
4. Inseparable design, soft margin
5. Kernels
 - a. Kernel definition (Hilbert spaces, inner product, positive semidefiniteness)
 - b. Kernels properties (addition, infinite sums)
 - c. Types of kernels (poly, exponential, gaussian)
6. Current state

History

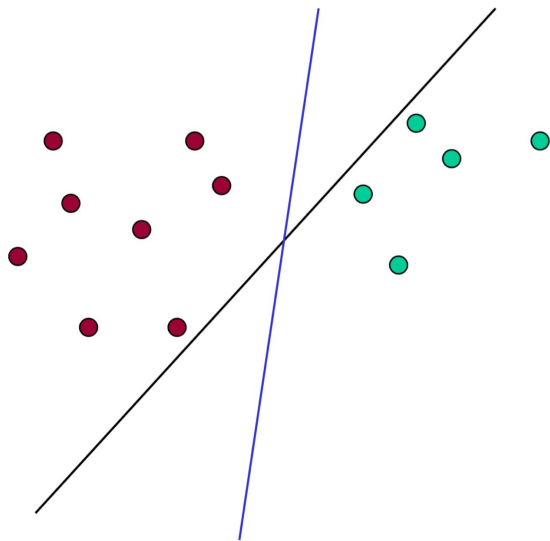


1963: SVM introduced by Soviet mathematicians
Vladimir Vapnik and Alexey Chervonenkis

1992: kernel trick (Vapnik, Boser, Guyon)

1995: soft margin (Vapnik, Cortes)

Motivation



Linear separable case

Many separating hyperplanes exist

Maximize width

Margin

$$y \in \{1, -1\}$$

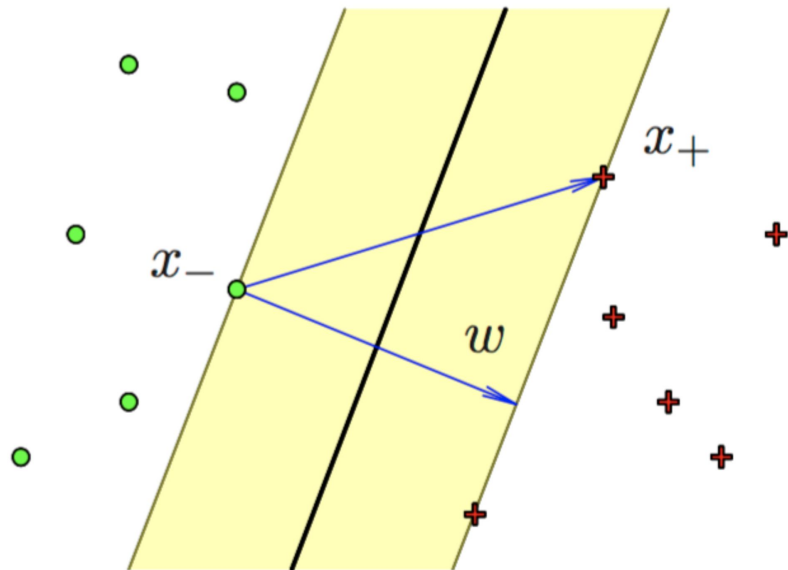
$$y_i = 1 : w^T x_i - c > 0$$

$$y_i = -1 : w^T x_i - c < 0$$

$$c_+(w) = \min_{y_i=1} (w^T x_i)$$

$$c_-(w) = \max_{y_i=-1} (w^T x_i)$$

$$\rho(w) = \frac{c_+(w) - c_-(w)}{2}$$



$$\rho \left(\frac{w_0}{||w_0||} \right) = \frac{1}{||w_0||}$$



Optimization problem



$$\begin{aligned} y_i = 1 & : w^T x_i - c > 0 \\ y_i = -1 & : w^T x_i - c < 0 \\ M_i & = y_i \cdot (w^T x_i - c) \end{aligned} \quad \begin{aligned} \rho(w) & = \frac{1}{||w||} \rightarrow \max_{w, c} \\ s.t. & \ y_i(w^T x_i - c) \geq 1 \end{aligned}$$

Convex problem!

$$L(w, c, \alpha) = \frac{1}{2} w^T w - \sum_i \alpha_i (y_i (w^T x_i - c) - 1)$$



Many of them are
zeros

Hinge loss



$$L(w, c, \alpha) = \frac{1}{2} w^T w - \sum_i \alpha_i (y_i (w^T x_i - c) - 1)$$

$$L^{\text{hinge}} = (1 - M)_+$$

$$L(w, c, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_i \alpha_i L_i^{\text{hinge}}$$

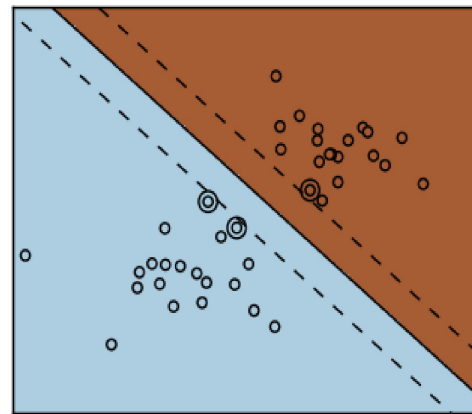
Inseparable case

Let our model mistake, but penalize that mistakes

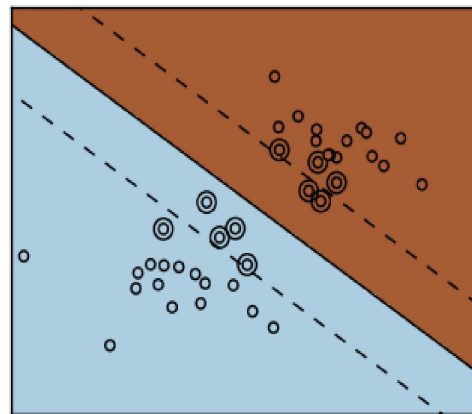
Implemented via margin slack variables

$$\begin{cases} \frac{1}{2} \langle w, w \rangle + C \sum_{i=1}^{\ell} \xi_i \rightarrow \min_{w, w_0, \xi}; \\ y_i (\langle w, x_i \rangle - w_0) \geq 1 - \xi_i, \quad i = 1, \dots, \ell; \\ \xi_i \geq 0, \quad i = 1, \dots, \ell. \end{cases}$$

Big C



Small C



Kernel trick

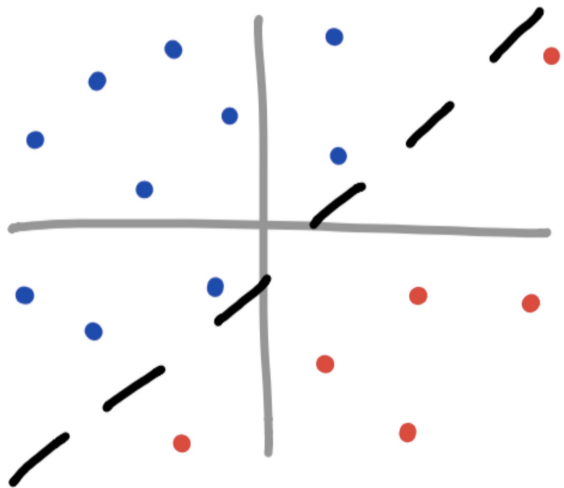


$$y_i = 1 : w^T x_i - c > 0$$
$$y_i = -1 : w^T x_i - c < 0$$

$$\begin{array}{l} x \mapsto \phi(x) \\ w \mapsto \phi(w) \end{array} \Rightarrow \langle w, x \rangle \mapsto \langle \phi(w), \phi(x) \rangle$$

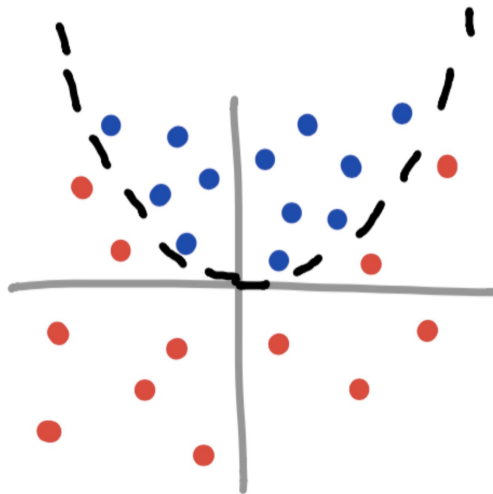
$$K(w, x) = \langle \phi(w), \phi(x) \rangle$$

Kernel types



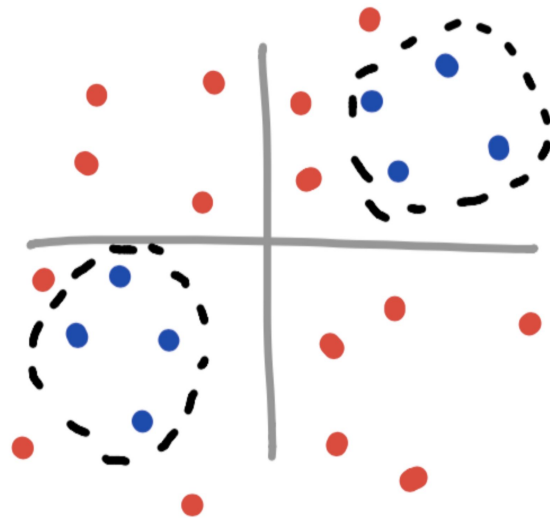
$$K(w, x) = \langle w, x \rangle$$

Linear



$$K(w, x) = (\gamma \langle w, x \rangle + r)^d$$

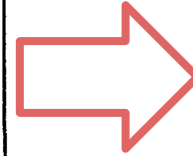
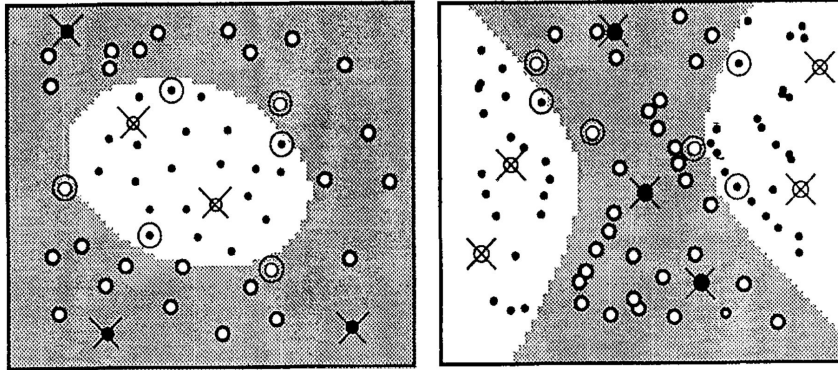
Polynomial



$$K(w, x) = e^{-\gamma \|w - x\|^2}$$

Gaussian
radial basis function

Current state



Principal Component Analysis

girafe
ai

02

Principal Component Analysis



$$x_1, \dots, x_n \rightarrow g_1, \dots, g_k, k \leq n$$

$$U : UU^T = I, G = XU$$

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

$$\|GU^T - X\| \rightarrow \min_{G,U} \text{ s.t. } \text{rank}(G) \leq k$$

Singular value decomposition



$$\|GU^T - X\| \rightarrow \min_{G,U} \text{ s.t. } \text{rank}(G) \leq k$$

$$X = V\Sigma U^T : \|GU^T - V\Sigma U^T\|_2 = \|G - V\Sigma\|_2$$

$$G = V\Sigma' : \|V\Sigma' - V\Sigma\|_2 = \|\Sigma' - \Sigma\|_2$$

$$\|A\|_2 = \sigma_{\max}(A) : \|\Sigma' - \Sigma\|_2 = \sigma_k(\Sigma) = \sigma_k(X)$$

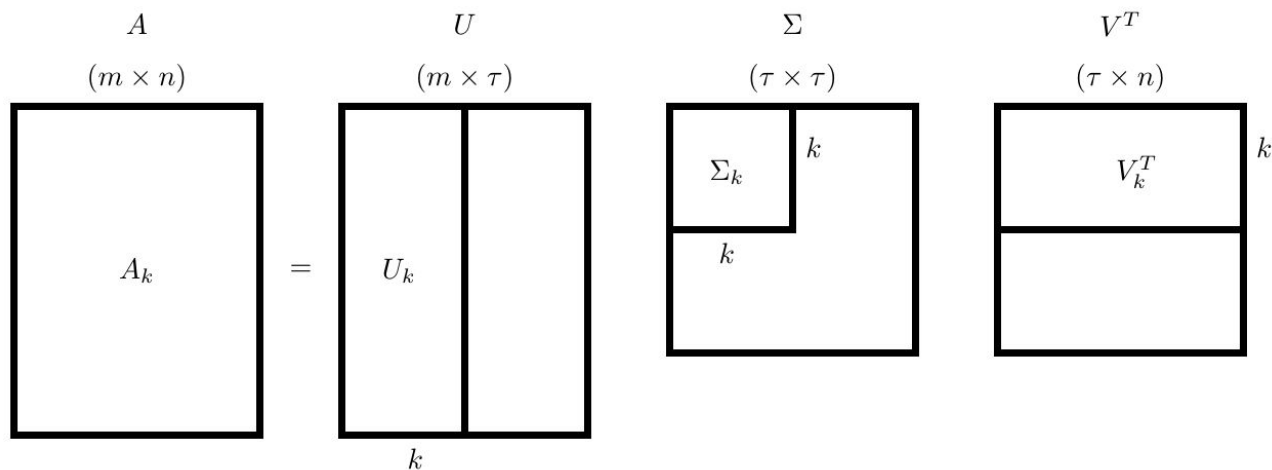
Eckart–Young–Mirsky theorem



Singular value decomposition

$$\|GU^T - X\| \rightarrow \min_{G,U} \text{ s.t. } \text{rank}(G) \leq k$$

$$X = V\Sigma U^T \quad \sigma_k(\Sigma) = \sigma_k(X)$$



Eckart-Young-Mirsky
theorem

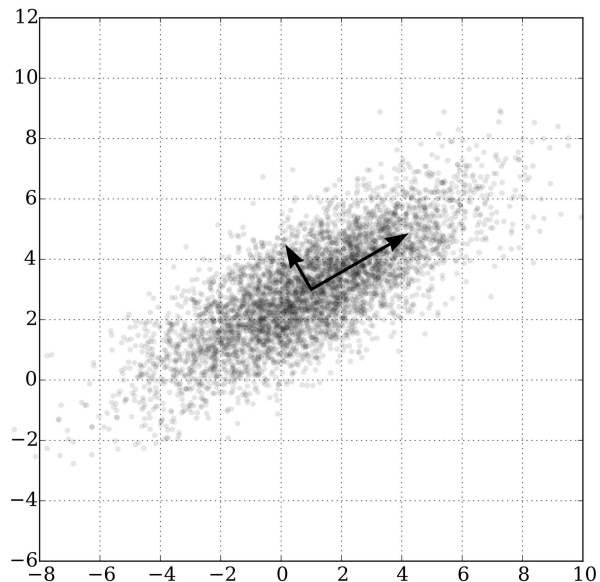
Another approach

Residual variance maximization

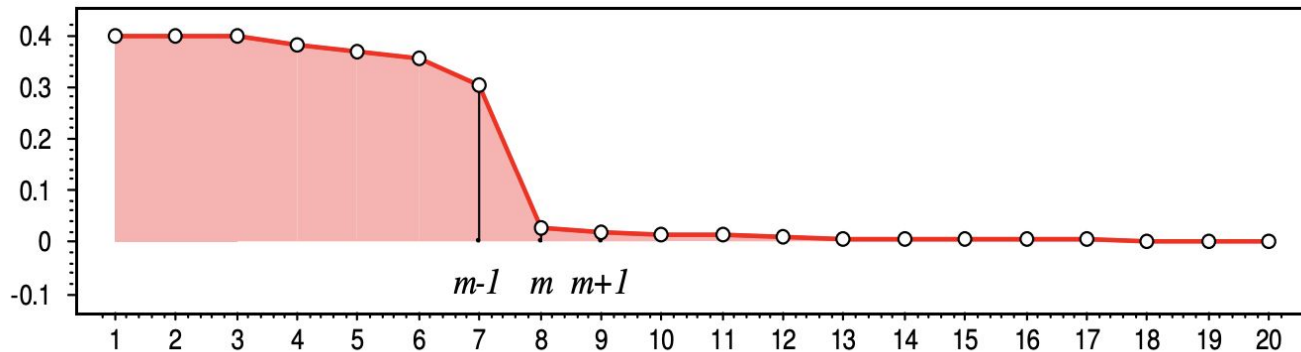
Take new basis vectors greedy

Same result for G and U

Always normalize data before PCA!!!



Dimensionality reduction



Get rid of low-variance components

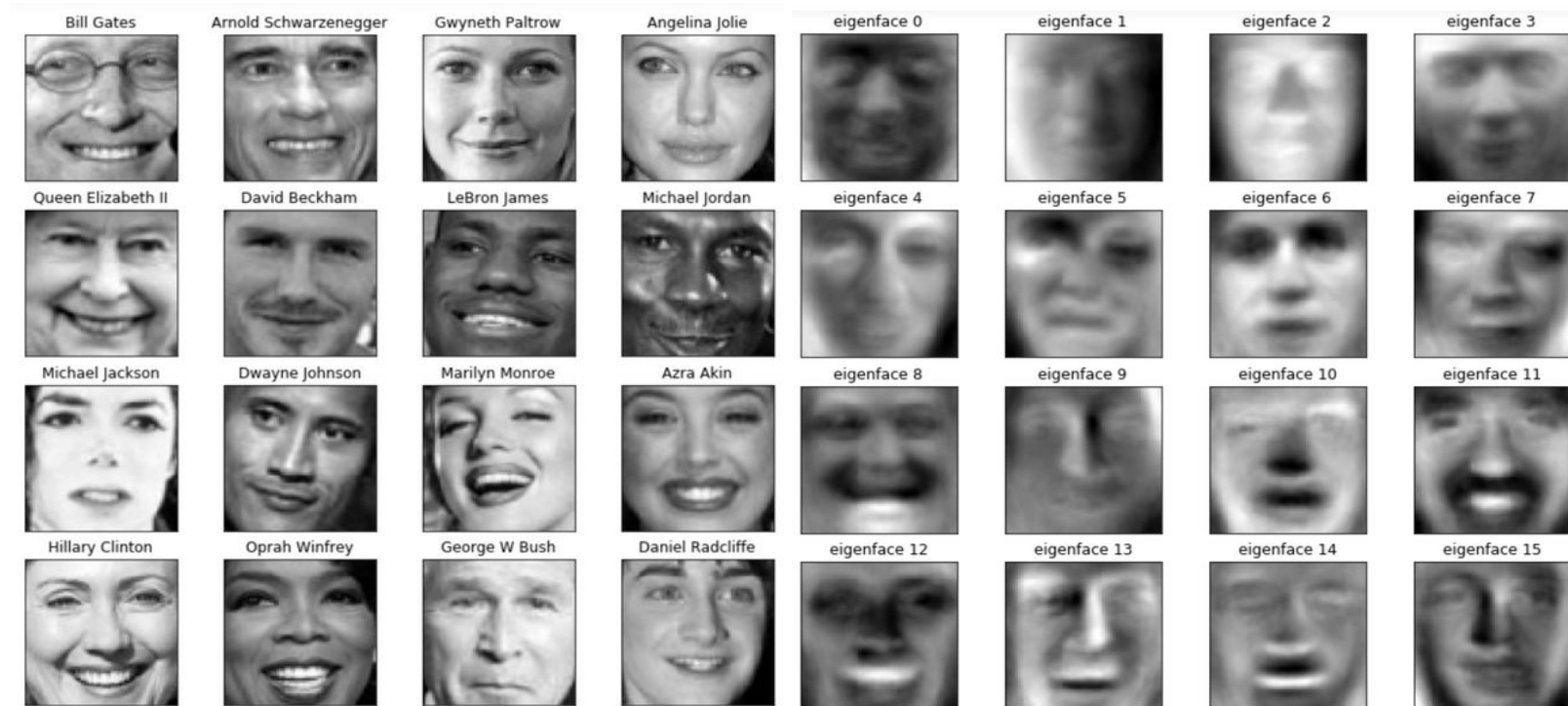
$$E_m = \frac{\|GU^T - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\lambda_1 + \dots + \lambda_n} \leq \varepsilon.$$

Dimensionality reduction



**Let's walk through
space...**

Dimensionality reduction



16 components

Dimensionality reduction



50 components

Dimensionality reduction



250 components

Revise



- Support Vector Machine (SVM)
 - History
 - Motivation
 - Solution for separable design
 - Inseparable design, soft margin
 - Kernels
- Dimensionality reduction and PCA
 - Problem statement
 - Eckart–Young theorem
 - Equivalent definitions

Next time

- Decision trees and thresholds

Thanks for attention!

Questions?



girafe
ai

