Principles of Machine Learning: Exercise 5

Alina Pollehn (3197257), Julian Litz (3362592), Manuel Hinz (3334548)
Felix Göhde (3336445), Felix Lehmann (3177181), Caspar Wiswesser (3221493)
Adrian Köring (3347785), Greta Günther (3326765), Linus Mallwitz (3327653)
Niklas Mueller-Goldingen (3363219), Jennifer Kroppen (2783393)

11.01.2024

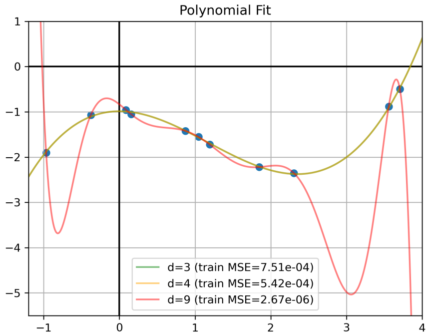
Exercise 5.1: Overview

1. Goal: Fitting a polynomial to noisy data i.e.via polynomial regression
2. First step: Transform inputs with feature map $\varphi(x) = [x^0, \dots, x^d]$ (aka Vandermonde-Matrix)

```
def vandermonde(x, degree=2): return np.vander(x, N=degree+1)
```

3. Second step: Estimate model weights: $\hat{w} = [\Phi\Phi^\intercal]^{-1}\Phi y$ (via numerically stable inversion (i.e. QR))
4. Third step: Inference with the fitted model: $\hat{f}(x) = \varphi(x)^\intercal \hat{w}$

Exercise 5.1.2: Results and Discussion

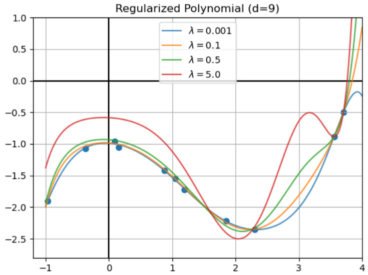


- The polynomial fit with degree $= 9$ results in a better MSE and is therefore the better model $\square$
- Now, the degree 3 model is *more reasonable* from (for example) the perspective of Occam's Razor.
- The point is: Judging overfitting from training data alone is questionable. We need validation data.
- In its abscence we can try **Leave-one-Out Cross-Validation** to quanitfy our intuition:

Degree	MSE
3	0.0033
4	0.0019
5	0.012
...	
9	151.06

Exercise 5.1.3: Adding regularization

We now investigate the effect of different λ on the solution to regularized least squares ($\hat{w} = [\Phi\Phi^\intercal + \lambda I]^{-1}\Phi y$):



- Good results for small $\lambda = 0.001$ within support, but shape towards $+\infty$ not as desired.
- Larger $\lambda = 0.1, 0.5$ yield better 'global' shape, but deviate more from x^3 within support.
- $\lambda = 5$ gives worse results, which makes sense, because we are adding a larger value, decreasing the impact of our gram matrix before inverting!

Exercise 5.2: Set-Up

- Lecture 07 showed that the dual least squares solution is given by $\hat{w} = \Phi[\Phi\Phi^\intercal]^{-1}y$
- After regularization this becomes $\hat{w} = \Phi[\Phi\Phi^\intercal + \lambda I]^{-1}y$
- We kernelize the expression

$$\hat{f}(x) = \phi^\intercal(x)\Phi[\Phi\Phi^\intercal + \lambda I]^{-1}y \tag{1}$$
$$\Rightarrow \hat{f}(x) = k(x)^\intercal[K + \lambda I]^{-1}y \tag{2}$$

- Initial choices for the model parameters where given: $\lambda = 0.5, b = 1, d = 3$

Exercise 5.2: Results

Influence of hyper-parameters on the estimated model:

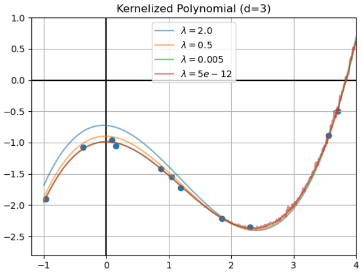


Figure: Kernelized Regression: Impact of various λ values

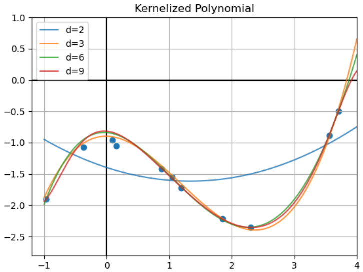


Figure: Kernelized Regression: Impact of various *degree* values

- λ regularizes the least-squares solution.
  - \+ Lower values lead to a more faithful fit
  - \- predictions contain more noise for very small values
- *degree* specifies the shape of the fitted function

Exercise 5.2: Results

Influence of hyper-parameters on the estimated model:

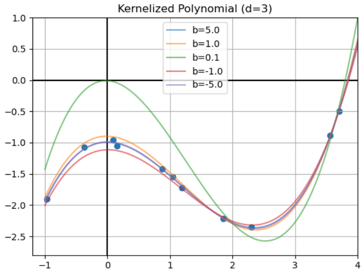


Figure: Kernelized Regression: Impact of various b values

- b also influences how faithful the model can stick to the measurements (mostly around $x = 0$)
- Small values lead to a small y-intercept for the predicted function
- Larger ones result in a better (overall) fit

Connection to GP

- Kernel Matrix with row/col for each data-point
- Prediction is weighted interpolation of training data

Exercise 5.3: Least squares SVMs for regression

- Adding to the lecture we can use SVMs for regression as well
- We want to build a least squares SVM regression model

$$\hat{f}(x) = \varphi(x)^\intercal \Phi \hat{\lambda} + \hat{b}$$

- Where

$$\begin{bmatrix} \hat{\lambda} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \Phi^\intercal \Phi + \frac{1}{C} I & 1 \\ 1^\intercal & 0 \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix}$$

- Which can easily be kernelized:

$$\begin{bmatrix} \hat{\lambda} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} K + \frac{1}{C} I & 1 \\ 1^\intercal & 0 \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix}$$

and

$$\hat{f}(x) = k(x)^\intercal \hat{\lambda} + \hat{b}$$

- Good results for a wider range of parameters (keeping the d fixed)
- Results differ more when changing the degree d

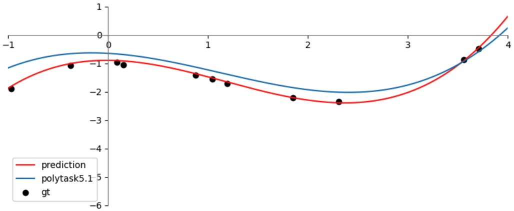


Figure: Polynomial fit for the given parameters

Other degrees

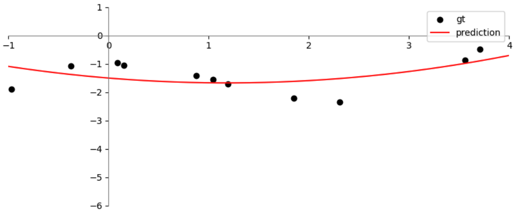


Figure: Polynomial fit for the given parameters $d = 2$

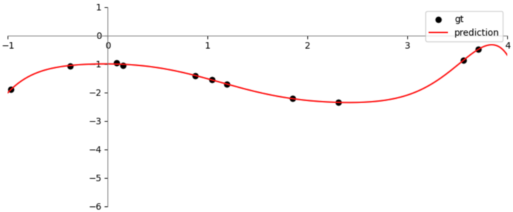


Figure: Polynomial fit for the given parameters $d = 9$

Exercise 5.4: Kernel SVM for binary classification

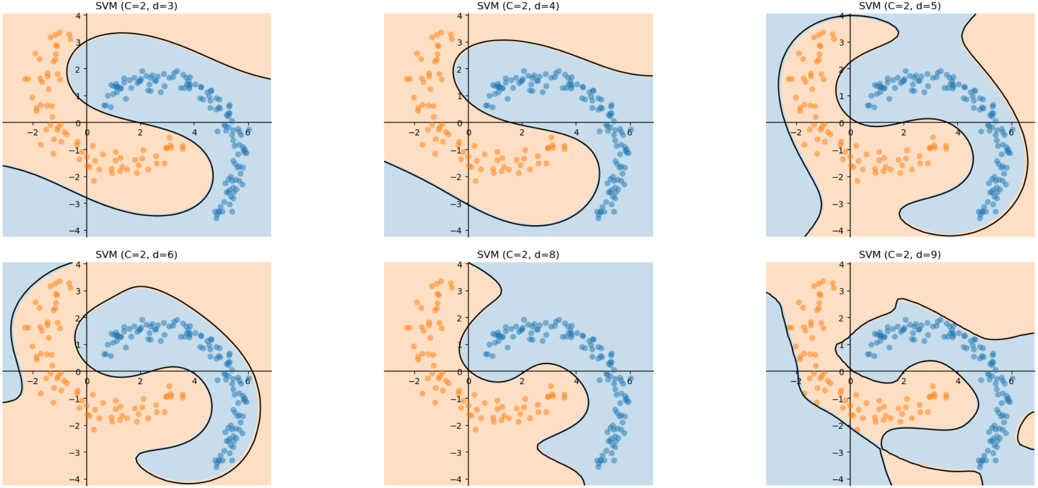
- Same math: $\begin{bmatrix} \hat{\boldsymbol{\lambda}} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \mathrm{K} + \frac{1}{C} I & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} y \\ 0 \end{bmatrix}$
- Different regression targets $y \in \{-1, 1\}$
- Polynomial Kernel $k(u, v) = (b + \boldsymbol{u}^\top \boldsymbol{v})^d$

```
K = (b + (X.T @ X))**d
I = 1/C * np.eye(len(K))
One = np.ones((len(K), 1))

M = np.block([[K + I, One],
              [One.T,   0]])
t = np.block([y, 0])

params, *_ = np.linalg.lstsq(M, t)
lam, bias = params[:-1], params[-1]
```

SVM Decision Boundary



Exercise 5.5: Minimum enclosing balls

- We already computed the minimal enclosing ball (MEB) for a given dataset in lecture 08:
  - Using Frank-Wolfe solve

$$\operatorname*{argmin}_{\mu} \mu^\intercal X^\intercal X \mu - \mu^\intercal z$$

  - where X denotes the dataset $X = [x_1, \ldots, x_n] \in \mathbb{R}^{m \times n}$ and $z = \mathrm{diag}[X^\intercal X]$
  - under the constraints $1^\intercal \mu = 1$ and $\mu \geq 0$
- given $\hat{\mu}$ we can than either compute the radius and the center of the ball

$$\hat{c} = X\hat{\mu} \text{ and } \hat{r} = \sqrt{\hat{\mu}^\intercal z - \hat{\mu}^\intercal X^\intercal X \hat{\mu}}$$

- which leads to a function

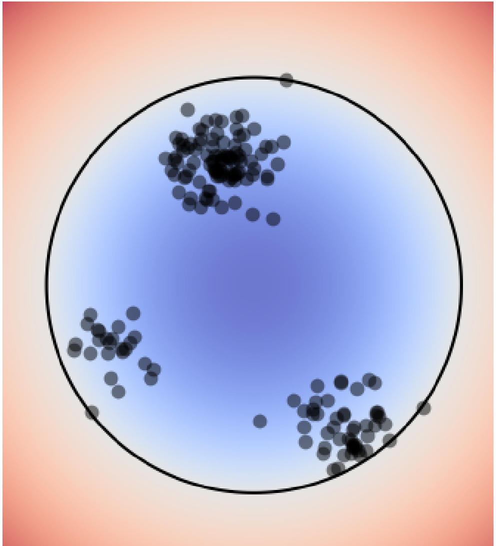
$$\chi_B(x) = \|x - \hat{c}\|^2 - \hat{r}^2$$

  which is negative for $x \notin B \cup \partial B$!
- we can also rewrite

$$\chi_B(x) = x^\intercal x - 2x^\intercal X\hat{\mu} + \hat{\mu}^\intercal X^\intercal X \hat{\mu} - \hat{\mu}^\intercal z + \hat{\mu}^\intercal X^\intercal X \hat{\mu}$$

Result and Frank-Wolfe-Implementation

```
def fwDualMEB(matX,vecZ,T=100):
    m, n = matX.shape
    vecM = np.ones(n)/n
    for t in range(T):
        beta = 2 / (t+2)
        vecG = 2 * matX.T @ \
            matX @ vecM - vecZ
        imin = np.argmin(vecG)
        vecM *= (1-beta)
        vecM[imin] += beta
    return vecM
```



Kernel minimum enclosing balls

- Using our second formulation

$$\chi_B(x) = x^\intercal x - 2x^\intercal X\hat{\mu} + \hat{\mu}^\intercal X^\intercal X\hat{\mu} - \hat{\mu}^\intercal z + \hat{\mu}^\intercal X^\intercal X + \hat{\mu}$$

  we can kernalize everything:
- We get:

$$\chi_B(x) = K(x,x) - 2\kappa^\intercal\hat{\mu} - \hat{\mu}^\intercal k + 2\hat{\mu}^\intercal K\hat{\mu}$$

- where $K(x,x) = \exp(0) = 1 \in \mathbb{R}$ and $k = 1 \in \mathbb{R}^n$, because $k_j = K(x_j, x_j) = \exp(0) = 1$.
- Using a Gaussian kernel:

$$k(u,v) = \exp(-\frac{1}{2\sigma^2}\|u-v\|^2)$$

  and the following Frank-Wolfe-Algorithm:

Frank-Wolfe-Algorithm for kernel minimum enclosing balls

- Solving the minimization problem problem

$$\begin{aligned} &\operatorname*{argmin}_{\mu} \mu^\intercal K \mu - \mu^\intercal k \\ = &\operatorname*{argmin}_{\mu} \mu^\intercal K \mu - \mu^\intercal 1 \end{aligned}$$

- under the constraints $1^\intercal \mu = 1$ and $\mu \geq 0$

```
def fwDualMEB2(K,k, T=100):
    m, n = K.shape
    vecM = np.ones(n)/n
    for t in range(T):
        beta = 2 / (t+2)
        vecG = K @ vecM - k
        imin = np.argmin(vecG)
        vecM *= (1-beta)
        vecM[imin] += beta
    return vecM
```

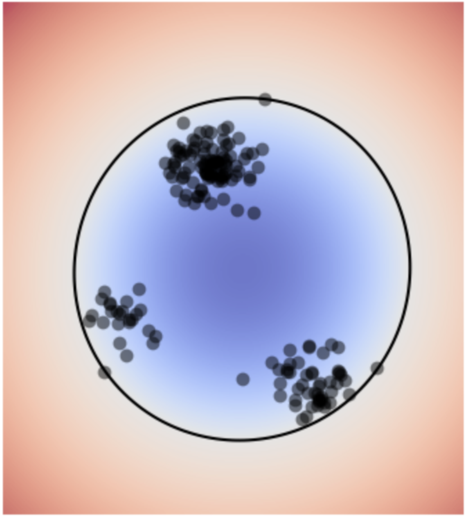


Figure: $\sigma = 4$

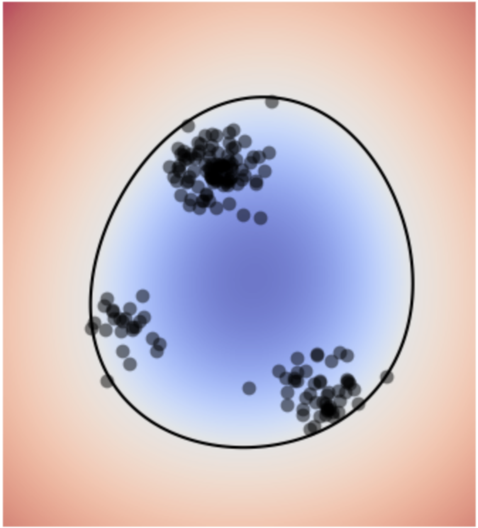


Figure: $\sigma = 2$

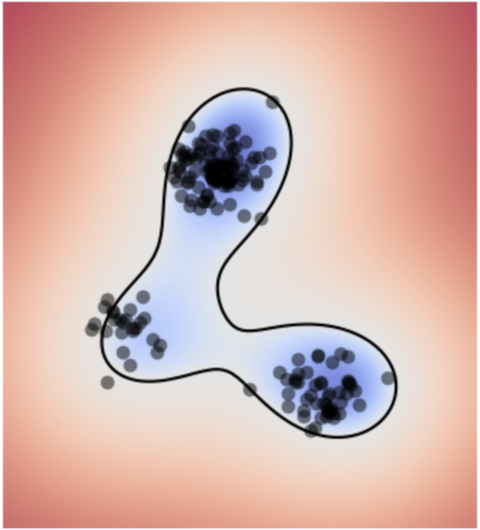


Figure: $\sigma = 1$

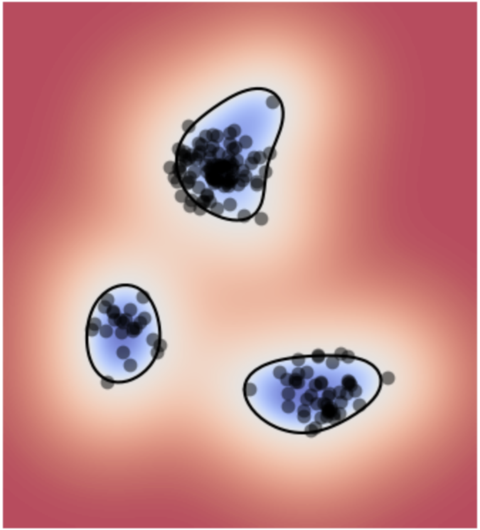


Figure: $\sigma = 0.5$