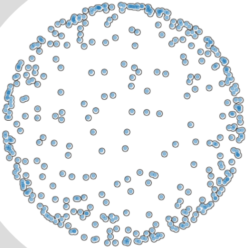
WiSe 2024/25

Machine Learning for Data Science

Lecture by G. Montavon

FREIE UNIVERSITÄT BERLIN



Lecture 7a | **Prediction**

Outline

Motivations

Least Square Regression

- Model and Objective

Connection to CCA

- Least square regression as a special case of CCA

Computational Aspects

- Problem of invertibility
- High-dimensions

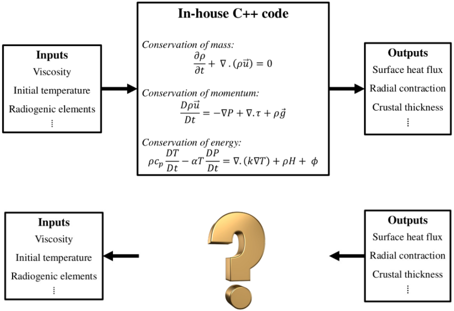
Support Vector Regression

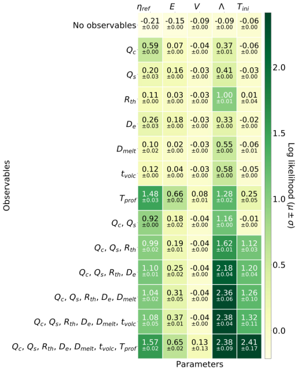
- Beyond least square regression
- Absolute deviations and ϵ-insensitivity
- SVR primal
- SVR dual

Motivations

Example: Space Science

- To what extent can physical parameters of (exo-)planets (e.g. viscosity) be inferred from observables [1].





Motivations

Example: Toxicity prediction

- Are particular substructures predictive of molecular toxicity, and can we predict toxicity from the molecular geometry [3].

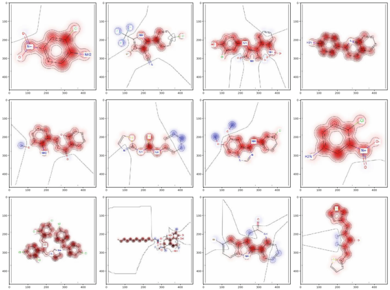


Image source: https://doi.org/10.1007/978-3-030-28954-6_18

Part 1

Least Squares Regression

Least Squares Regression

Idea:

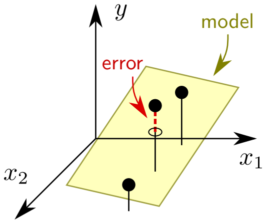
- Build a model predicting y from $\boldsymbol{x}$:

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

- Find model parameters so that the prediction errors are minimized, e.g. least square error averaged over all instances:

$$\min_{\boldsymbol{w},b} \mathrm{E}[(\boldsymbol{w}^\top \boldsymbol{x} + b - y)^2]$$

- The error of the model can either be of interest by itself (quantifying predictability), or the learned parameters can be inspected for further insights.



Simplification for Centered Data

Proposition

Assume we wish to train a linear model

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

on data that has been centered as a first step, i.e. $\mathrm{E}[\boldsymbol{x}] = \boldsymbol{0}$ *and* $\mathrm{E}[y] = 0$. *Then, we can show that*

$$\arg\max_b \mathrm{E}[(f(\boldsymbol{x}) - y)^2] = 0$$

i.e. it is always best to set the bias to zero.

Proof:

$$\frac{\partial}{\partial b} \mathrm{E}[(f(\boldsymbol{x}) - y)^2] = 2\mathrm{E}[(\boldsymbol{w}^\top \boldsymbol{x} + b - y)] \stackrel{\text{def}}{=} 0 \Rightarrow b = 0$$

$\square$

Implication:

- Without loss of accuracy, one can simplify the least square regression problem as finding a homogeneous linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ that minimizes the mean square error.

Linear Regression

- Recall that we consider prediction functions of the type $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ and we would like to minimize the mean square error. The latter can be developed as:

$$\begin{aligned}
\mathcal{E}(\boldsymbol{w}) &= \mathrm{E}[(f(\boldsymbol{x}) - y)^2] \\
&= \mathrm{E}[(\boldsymbol{w}^\top \boldsymbol{x} - y)^2] \\
&= \mathrm{E}[\boldsymbol{w}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{x} y + y^2] \\
&= \boldsymbol{w}^\top C_{xx} \boldsymbol{w} - 2\boldsymbol{w}^\top C_{xy} + C_{yy}
\end{aligned}$$

- Observing that C_{xx} is positive semi-definite, minimizing $\mathcal{E}(\boldsymbol{w})$ is a convex optimization problem and the solution must necessarily have gradient zero, i.e. $\nabla\mathcal{E}(\boldsymbol{w}) = 2C_{xx}\boldsymbol{w} - 2C_{xy} = 0$. Solving for $\boldsymbol{w}$ gives us the optimal model:

$$\boxed{\boldsymbol{w} = C_{xx}^{-1} C_{xy}}$$

- Injecting this solution into the objective, we get the mean square error at the optimum:

$$\mathcal{E} = C_{yy} - C_{yx} C_{xx}^{-1} C_{xy}$$

Example of a Linear Model

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		1 B	Std. Error	3 Beta	t	Sig. 2	Lower Bound	Upper Bound
1	(Constant)	-3263.586	1059.163		-3.081	.002	-5344.360	-1182.812
	Sex	509.271	139.565	.126	3.649	.000	235.089	783.454
	Age at Survey Completion (Years)	114.658	12.452	.322	9.208	.000	90.196	139.121
	Average Consumption of Alcoholic Beverages per Week	50.386	10.275	.192	4.904	.000	30.201	70.572
	Average Consumption of Cigarettes per Day	139.414	17.384	.311	8.020	.000	105.263	173.565
	Average Hours of Exercise per Week	-271.270	36.300	-.281	-7.473	.000	-342.584	-199.956

a. Dependent Variable: Total Health Care Costs Declared over 2020

Image source: `https://www.spss-tutorials.com/spss-multiple-linear-regression-example/`

Remarks:

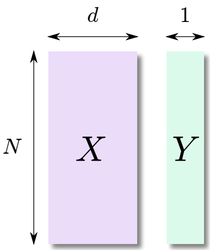
- Model weights can be interpreted as the amount by which the output would increase if changing the input variable by $+1$. To assess the relative importance of variables, it is necessary to standardize the input features.
- Unless the input data are decorrelated (C_{xx} diagonal), it is not impossible that input features that positively correlate with the output contribute *negatively* in the linear model. $\rightarrow$ one must be careful with interpreting the weights of a linear model.

Part 2 **Connection to CCA**

Connection to CCA

- Recall that canonical correlation analysis aims to find a projection of two modalities that maximizes their correlation, and that such correlation can be expressed in terms of auto- and cross-covariances as:

$$\rho = \frac{\boldsymbol{w}_x^\top C_{xy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^\top C_{xx} \boldsymbol{w}_x} \sqrt{\boldsymbol{w}_y^\top C_{yy} \boldsymbol{w}_y}}$$



- Because in the case of linear regression the second modality is univariate, there is no direction to be found in the second modality (we can set manually $w_y = 1$), and the CCA objective simplifies to:

$$\rho = \frac{\boldsymbol{w}_x^\top C_{xy}}{\sqrt{\boldsymbol{w}_x^\top C_{xx} \boldsymbol{w}_x C_{yy}}}$$

- This can be reformulated as a constrained optimization problem:

$$\max_{\boldsymbol{w}} \boldsymbol{w}_x^\top C_{xy} \qquad \text{s.t.} \quad \boldsymbol{w}_x^\top C_{xx} \boldsymbol{w}_x C_{yy} = 1$$

Connection to CCA: Deriving the Weights

- We would to solve the constrained optimization problem:

$$\max_{\boldsymbol{w}} \boldsymbol{w}_x^\top C_{xy} \qquad \text{s.t.} \quad \boldsymbol{w}_x^\top C_{xx} \boldsymbol{w}_x C_{yy} = 1$$

- Using the framework of Lagrange multipliers, we get:

$$\mathcal{L}(\boldsymbol{w}_x, \lambda) = \boldsymbol{w}_x^\top C_{xy} + \frac{1}{2}\lambda(1 - \boldsymbol{w}_x^\top C_{xx} \boldsymbol{w}_x C_{yy})$$

and

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_x} = C_{xy} - \lambda C_{xx} \boldsymbol{w}_x C_{yy} \stackrel{!}{=} 0 \quad \Rightarrow \quad \boldsymbol{w}_x = \lambda^{-1} C_{xx}^{-1} C_{xy} C_{yy}^{-1}$$

and setting λ in a way that the constraint is satisfied, we get:

$$\boldsymbol{w}_x = \frac{C_{xx}^{-1} C_{xy}}{\sqrt{C_{yx} C_{xx}^{-1} C_{xy} C_{yy}}}.$$

- This is exactly the same direction as the weight of the linear model ($\boldsymbol{w} = C_{xx}^{-1} C_{xy}$).

Connection to CCA: Deriving the Objective

- Recall that the parameter that optimizes our CCA objective is given by:

$$\boldsymbol{w}_x = \frac{C_{xx}^{-1} C_{xy}}{\sqrt{C_{yx} C_{xx}^{-1} C_{xy} C_{yy}}}$$

- Evaluating the CCA objective with this solution gives the correlation coefficient

$$\rho = \boldsymbol{w}_x^\top C_{xy} = \sqrt{C_{yx} C_{xx}^{-1} C_{xy} C_{yy}^{-1}}.$$

  From this correlation coefficient, one can measure the explained variance $C_{yy}\rho^2 = C_{yx} C_{xx}^{-1} C_{xy}$.

- Compare with the error of our linear model that was given by:

$$\mathcal{E} = C_{yy} - C_{yx} C_{xx}^{-1} C_{xy}$$

- The objectives of CCA and least square regression are therefore related as $C_{yy} = C_{yy}\rho^2 + \mathcal{E}$, i.e. together they form a decomposition of the total variance (i.e. maximizing the correlation using CCA or minimizing the mean square error achieve both the same objective!).

Part 3

Computational Aspects

Back to Linear Regression ...

Recall:

- When considering prediction functions of the type $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, and minimizing the prediction error

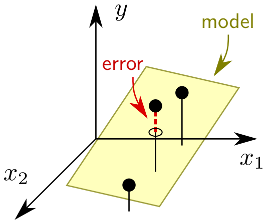
  $$\mathcal{E}(\boldsymbol{w}) = \mathrm{E}[(f(\boldsymbol{x}) - y)^2].$$

  we get the optimal model model:

  $$\boldsymbol{w} = C_{xx}^{-1} C_{xy}$$

  and its error is:

  $$\mathcal{E} = C_{yy} - C_{yx} C_{xx}^{-1} C_{xy}$$



Question:

- What if C_{xx} is not invertible?

Dealing with Non-Invertible C_{xx}

Observation:

- C_{xx} non-invertible corresponds to the case where the data only spans a subspace of $\mathbb{R}^d$. In that case, there are infinitely many possible solutions to the regression problem.

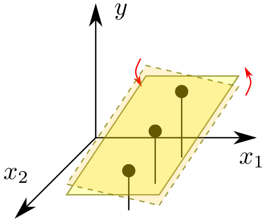
Idea:

- (Virtually) add some small uncorrelated noise $\boldsymbol{n}$ to the data.

$$\begin{aligned} C_{xx}^{\text{new}} &= \mathrm{Cov}(\boldsymbol{x}+\boldsymbol{n}, \boldsymbol{x}+\boldsymbol{n}) \\ &= \mathrm{Cov}(\boldsymbol{x},\boldsymbol{x}) + 2\mathrm{Cov}(\boldsymbol{x},\boldsymbol{n}) + \mathrm{Cov}(\boldsymbol{n},\boldsymbol{n}) \\ &= C_{xx} + \sigma_n^2 I \end{aligned}$$

  This corresponds to adding a diagonal term to the covariance matrix.

- This makes the covariance matrix invertible and favors solutions that are flat on directions that are orthogonal to the data.



Dealing with Non-Invertible C_{xx}

Implication on the model error:

- Let us virtually inject the uncorrelated noise in the data, obtain the resulting robustified covariance matrix $C_{xx}^{\text{new}} = C_{xx} + \sigma_n^2 I$, and also observe that the covariance matrices C_{xy}, C_{yy} are not affected by that noise.
- Using the new covariances in the error function yields:

$$\begin{aligned}
\mathcal{E}^{\text{new}} &= C_{yy} - C_{yx}\big(C_{xx}^{\text{new}}\big)^{-1} C_{xy} \\
&= C_{yy} - C_{yx}(C_{xx} + \sigma_n^2 I)^{-1} C_{xy} \\
&= C_{yy} - C_{yx}(\textstyle\sum_{j=1}^{d} u_j u_j^\top (\lambda_j + \sigma_n^2))^{-1} C_{xy} \\
&= C_{yy} - \textstyle\sum_j (C_{yx} u_j)^2 \frac{1}{\lambda_j + \sigma_n^2}
\end{aligned}$$

The larger the noise σ_n^2, the higher the error (and the lower the correlation).
$\rightarrow$ Just use σ_n large enough to be able to invert C_{xx}.

Note:

- This is not the whole picture yet, a higher value of σ_n^2 can be beneficial for *reproducibility* of results ($\rightarrow$ next week).

Dealing with High Dimensions

- When d is large, computing and inverting the matrix C_{xx} can be prohibitively expensive.
- Let us start from the original formulation of the regression objective

$$\begin{aligned}\mathcal{E}(\boldsymbol{w}) &= \mathrm{E}[(\boldsymbol{w}^\top \boldsymbol{x} - y)^2] \\ &= \boldsymbol{w}^\top C_{xx} \boldsymbol{w} - 2\boldsymbol{w}^\top C_{xy} + C_{yy}\end{aligned}$$

- We now assume that an optimal $\boldsymbol{w}$ is found in the span of the data, and express $\boldsymbol{w}$ as $\boldsymbol{w} = X\boldsymbol{\alpha}$, a linear combination of the data, where X is a matrix of size $d \times N$ containing the centered data:

$$\begin{aligned}\mathcal{E}(\boldsymbol{\alpha}) &= \overbrace{\boldsymbol{\alpha}^\top X^\top}^{\boldsymbol{w}^\top} C_{xx} \overbrace{X\boldsymbol{\alpha}}^{\boldsymbol{w}} - 2\,\overbrace{\boldsymbol{\alpha}^\top X^\top}^{\boldsymbol{w}^\top} C_{xy} + C_{yy} \\ &= \frac{1}{N}\boldsymbol{\alpha}^\top \underbrace{X^\top X X^\top X}_{Q_x^2} \boldsymbol{\alpha} - \frac{2}{N}\boldsymbol{\alpha}^\top \underbrace{X^\top X}_{Q_x} Y + C_{yy}\end{aligned}$$

- No need to compute matrices of size $d \times d$. Matrices Q_x^2 and Q_x are of size $N \times N$.

Dealing with High Dimensions

- We have found that the least square error can be expressed as

$$\mathcal{E} = \frac{1}{N}\boldsymbol{\alpha}^\top Q_x^2 \boldsymbol{\alpha} - \frac{2}{N}\boldsymbol{\alpha}^\top Q_x Y + C_{yy}$$

- Observing that $\mathcal{E}(\boldsymbol{\alpha})$ is convex, we find the solution where the gradient is zero:

$$\nabla \mathcal{E}(\boldsymbol{\alpha}) = \frac{2}{N} Q_x^2 \boldsymbol{\alpha} - \frac{2}{N} Q_x Y \stackrel{!}{=} 0$$

  and we find the solution

$$\boldsymbol{\alpha} = (Q_x^2)^{-1} Q_x Y = Q_x^{-1} Y$$

  Therefore, we only need to invert a matrix of size $N \times N$.

- From this solution, one can recover the original weight parameter as:

$$\boldsymbol{w} = X\boldsymbol{\alpha}$$

Part 4

Regression with Outliers

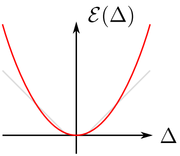
Regression with Outliers

Motivations:

- Square errors may be very large for data points whose outputs strongly deviate from that of other data points.
- Points with large prediction errors may be better treated as outliers (for which we accept the model to be inaccurate), so that the model can focus on the non-outlier part of the data.
- Reduced sensitivity to outliers can be achieved by considering absolute deviations instead of square errors, and invariance to small noise can be maintained by introducing a small slack ϵ.

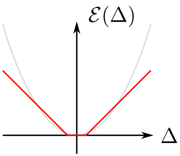
squared error

$$\mathcal{E}(\Delta) = \Delta^2$$



ϵ-insensitive absolute deviation

$$\mathcal{E}(\Delta) = \max(0, |\Delta| - \epsilon)$$



Regression with Outliers

- Consider the original least square regression problem

$$\min_{\boldsymbol{w}} \mathrm{E}\big[\underbrace{(\boldsymbol{w}^\top x - y)}_{\Delta}{}^2\big]$$

- To reduce exposure to outliers, we can replace the square error by the (ϵ-insensitive) absolute deviation.

$$\min_{\boldsymbol{w},b} \mathrm{E}\big[\max(0, |\underbrace{\boldsymbol{w}^\top x + b - y}_{\Delta}| - \epsilon)\big]$$

  (note that we need to reintroduce the bias, because we are no longer guaranteed that a solution without bias is optimal).

- There may be be multiple solutions that solve the problem exactly (especially when $d > N$), hence, we introduce a small penalty term $\lambda \|\boldsymbol{w}\|^2$ that favors the flattest solution:

$$\min_{\boldsymbol{w},b} \mathrm{E}\big[\max(0, |\underbrace{\boldsymbol{w}^\top x + b - y}_{\Delta}| - \epsilon)\big] + \lambda \|\boldsymbol{w}\|^2$$

Support Vector Regression

An algorithm that implements these ideas is Support Vector Regression [2, 4]. It takes the form of a quadratic optimization problem with linear constraints.

Support Vector Regression (Primal):

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^\star} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^\star)$$

$$\text{s.t.} \qquad \begin{aligned} &\forall_{i=1}^{N}: \ \boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i \leq \epsilon + \xi_i \\ &\forall_{i=1}^{N}: \ y_i - \boldsymbol{w}^\top \boldsymbol{x}_i + b \leq \epsilon + \xi_i^\star \\ &\qquad\qquad \xi_i, \xi_i^\star \geq 0 \end{aligned}$$

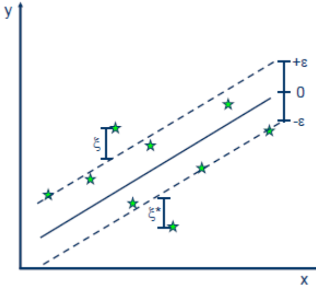


Image source: `https://www.saedsayad.com/support_vector_machine_reg.htm`

SVR: Primal vs. Dual

Primal:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^\star} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^N(\xi_i + \xi_i^\star)$$

$$\text{s.t.} \qquad \forall_{i=1}^N: \ \boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i \leq \epsilon + \xi_i$$

$$\forall_{i=1}^N: \ y_i - \boldsymbol{w}^\top \boldsymbol{x}_i + b \leq \epsilon + \xi_i^\star$$

$$\xi_i, \xi_i^\star \geq 0$$

Dual: (adapted from [4])

$$\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^\star} \quad -\frac{1}{2}\sum_{ij}(\alpha_i - \alpha_i^\star)(\alpha_j - \alpha_j^\star)\boldsymbol{x}_i^\top \boldsymbol{x}_j - \epsilon\sum_i(\alpha_i + \alpha_i^\star) + \sum_i y_i(\alpha_i - \alpha_i^\star)$$

$$\text{subject to} \quad \sum_i(\alpha_i - \alpha_i^\star) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^\star \in [0, C]$$

Like for the one-class SVM, the weights and bias of the primal problem can be recovered from the dual solution and the KKT conditions (cf. [4]).

Support Vector Regression

Advantages:

- Enables the model to be robust to outliers in the data and also to small noise perturbations ($\rightarrow$ matches well data encountered in practice).
- The SVR primal and dual both take the form of a quadratic optimization problem with linear constraints ($\rightarrow$ one can use standard solvers and always finds the true optimum of the objective).

Disadvantages:

- Unlike least squares regression, SVR has no closed form solution.
- Unlike least squares regression, there is no simple relation between the SVR objective and statistical measures such as correlation or variance.

Summary

Summary

- Regression is a very common data analysis, that gives us insights into the relation between a set of input variables and an output variable of interest, e.g. what correlation can we achieve between these variables, how the output variable responds to the different input variables.
- The most common formulation is the least square regression. It has an analytical solution and connections to the more general canonical correlation analysis.
- Least square error is not robust to outliers. In presence of outliers, it is better to consider absolute deviations. This can be addressed within the framework of support vector regression.

Further Topics

Ridge Regression:

- Implement a preference for models that are flatter, even at the cost of incurring additional prediction errors on the available data.
- This can be useful for improving the reproducibility of insights extracted from the model ($\rightarrow$ Lecture 8).

Lasso Regression:

- Implement a preference for models that respond only to a few input variables.
- This can be useful for improving the interpretability of the model and its predictions ($\rightarrow$ Lecture 8)

Nonlinear Regression:

- Replace the linear mapping between input and output by a nonlinear mapping, e.g. quadratic discriminants, kernel ridge regression, kernel SVR, deep neural networks.
- This enhances the representation power of regression models and enables reducing the prediction error / strengthening correlations ($\rightarrow$ Lectures 10-12).

References

S. Agarwal, N. Tosi, P. Kessel, S. Padovan, D. Breuer, and G. Montavon.
Toward constraining mars' thermal evolution using machine learning.
Earth and Space Science, 8(4), Apr. 2021.

H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik.
Support vector regression machines.
In *NIPS*, pages 155–161. MIT Press, 1996.

K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, and T. Unterthiner.
Interpretable deep learning in drug discovery.
In *Explainable AI*, volume 11700 of *Lecture Notes in Computer Science*, pages 331–345. Springer, 2019.

A. J. Smola and B. Schölkopf.
A tutorial on support vector regression.
Stat. Comput., 14(3):199–222, 2004.