Lectures 5 & 6: Linear Regression

Isabel Valera

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

03.05.2021 & 05.05.2021

Outline

1. Bibliography
2. Least squares regression
3. Basis functions
4. Bias-variance
5. Gauss-Markov-Theorem
6. Ridge Regression
7. Lasso
8. Summary
9. Appendix

Main references

- Bishop - Chapter 3
- ESL - Chapter 3

Outline

1. Bibliography
2. Least squares regression
3. Basis functions
4. Bias-variance
5. Gauss-Markov-Theorem
6. Ridge Regression
7. Lasso
8. Summary
9. Appendix

Regression setting

Regression setting: We consider problems where the output (target) variable is real-valued, i.e., $\mathcal{Y} = \mathbb{R}$. Moreover, we assume we have access to training data $(X_i, Y_i)_{i=1}^n$, which is an i.i.d. sample from the probability measure P on $\mathcal{X} \times \mathcal{Y}$. Note that here we treat each observation (X_i, Y_i) in the training dataset as a random variable.

Goal: Learn a mapping function $f^*(X)$ that minimizes the risk $R(f) = \mathbb{E}\big[L(Y, f(X))\big]$ with $f(X) \in \mathcal{F}$.

Standard loss is the squared loss, i.e., $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$, leading to a **least squares regression** problem.

Linear regression considers a family of regression functions $\mathcal{F}$ that are linear, i.e., it takes the form $\Big\{ \langle \mathbf{w}, \mathbf{x} \rangle + b, \ \text{with } \mathbf{x}, \mathbf{w} \in \mathbb{R}^d, \ b \in \mathbb{R} \Big\}$.

Illustration of Linear LSR

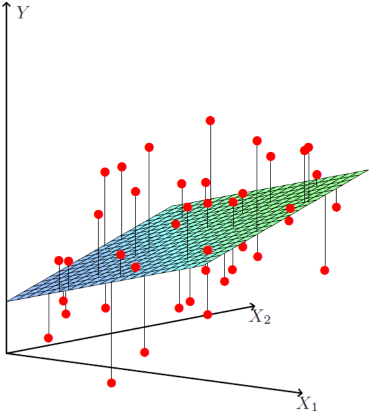


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Figure: Figure from ESL book.

Least squares regression

Least squares regression (LSR) considers the **Risk of squared loss**, i.e.,

$$\mathbb{E}[(Y - f(X))^2] = \mathbb{E}\left[\mathbb{E}\left[(Y - f(X))^2 \,|\, X\right]\right].$$

The optimal solution to LSR, i.e., the **Bayes optimal function**, is given by:

$$f^*(\mathbf{x}) = \mathbb{E}\left[Y|X = \mathbf{x}\right].$$

Definition (Least Squares Regression)

Given a training sample $D_n = (\mathbf{x}_i, y_i)_{i=1}^n$ with $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$, and a function space $\mathcal{F}$ we define **least squares regression** solution as

$$f_n^* = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2.$$

Linear LSR

Linear least squares regression assumes a linear function class, i.e,

$$\mathcal{F} = \left\{ f \;\middle|\; f(x) = \sum_{i=1}^{d} w_i x_i + b = \langle \mathbf{w}, \mathbf{x} \rangle + b, \quad w \in \mathbb{R}^d,\ b \in \mathbb{R} \right\}.$$

Notation:

- summarize the outputs $(y_i)_{i=1}^n$ into a column vector $\mathbf{Y} \in \mathbb{R}^n$ and the input vectors $(\mathbf{x}_i)_{i=1}^n$ into a feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}.$$

- $\mathbf{w}$ is the **weight vector** and b is the **bias term** (a.k.a. intercept).

Note: During the course, we consider vectors as column vectors.

Observation: One may alternatively include the bias term in the weight vector by adding an additional column with all values equal to one to the feature matrix.

Matrix form

Add extra dimension to input vector to integrate constant term in the (now) $(d+1)-$dimensional weight vector,

$$\mathbf{x}'_i = (1, x_{i1}, \dots, x_{id}) \qquad \text{or} \qquad x'_{i(0)} = 1,\ \forall i.$$

An affine function is characterized by the weight vector w,

$$\mathbf{w} \in \mathbb{R}^{d+1},\ f(\mathbf{x}'_i) = \langle \mathbf{w}, \mathbf{x}'_i \rangle = \sum_{j=0}^{d} w_j x'_{ij} = \sum_{j=1}^{d} w_j x_{ij} + w_0.$$

Linear least squares regression:

$$\mathbf{w}_n = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^{d+1}} \quad \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}'_i, \mathbf{w} \rangle)^2$$

Note: we will make the constant $w_{n0} = b$ implicit along the lecture by assuming that it is already included in the d dimensions/features of the observed features.

Solution of Linear LSR

Proposition

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$. *The solution* $\mathbf{w}_n$ *of linear least squares regression is given by*

$$\mathbf{w}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

where the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ *exists if* $\mathbf{X}$ *has rank* d.

If $\mathbf{X}$ *has rank lower than* d, *then* $(\mathbf{X}^T\mathbf{X})^{-1}$ *has to be understood in the sense of a generalized inverse. In this case the solution is not unique but if* $\mathbf{w}_n^1, \mathbf{w}_n^2$ *are two different solutions, then their predictions agree on the training data*

$$f_{\mathbf{w}_n^1}(\mathbf{x}_i) = \left\langle \mathbf{w}_n^1, \mathbf{x}_i \right\rangle = \left\langle \mathbf{w}_n^2, \mathbf{x}_i \right\rangle = f_{\mathbf{w}_n^2}(\mathbf{x}_i), \quad \forall\, i = 1, \ldots, n.$$

Note: See 9 Appendix for details on the generalized inverse.

Linear LSR solution Proof

Proof: Objective function of the optimization problem with $\mathbf{w} \in \mathbb{R}^d$,

$$R_{LLSR}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2 .$$

Taking the derivative with respect to $\mathbf{w}$,

$$\nabla_{\mathbf{w}} R_{LLSR} = -\frac{2}{n}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{w}).$$

The necessary condition for an extremum of R_{LLSR} is therefore

$$\frac{2}{n}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{w}) = 0 \quad \rightarrow \quad \mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})\mathbf{w} \rightarrow \mathbf{w}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

The solution $\mathbf{w}_n$ is unique, if the Hessian of $\mathbf{X}^T\mathbf{X}$ is positive-definite, which occurs if $\mathbf{X}$ has rank d. If $\mathbf{X}$ has rank smaller than d, then the solution is not unique.

Outline

1. Bibliography
2. Least squares regression
3. **Basis functions**
4. Bias-variance
5. Gauss-Markov-Theorem
6. Ridge Regression
7. Lasso
8. Summary
9. Appendix

Basis functions

Basis functions/Feature maps: Can be used to perform linear regression when the dependence between the features and outputs is non-linear but we have prior knowledge on the "non-linearity".

- The idea is to map the input $x \to \phi(x)$ into a new space in which the relationship between the new features X' and the outcome Y is linear.
- Examples:
  $\mathcal{X} = \mathbb{R}, \quad \text{then} \quad \phi(x) = 1, x, x^2, x^3, \ldots$ (polynomials),
  $\mathcal{X} = [0, 2\pi], \quad \text{then} \quad \phi(x) = \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots$
  (Fourier basis).

Assuming a (fixed) pre-defined set of M **basis functions**, $\phi_i : \mathbb{R}^d \to \mathbb{R}$, we can define the function space as:

$$\mathcal{F} = \left\{ f : \mathbb{R}^M \to \mathbb{R},\ f(\mathbf{x}) = \sum_{i=1}^{M} w_i\, \phi_i(\mathbf{x}) \,|\, \mathbf{w} \in \mathbb{R}^M \right\},$$

which remains linear with respect to the weight vector $\mathbf{w}$.

LSR (generalization)

Generalized feature matrix: $\Phi \in \mathbb{R}^{n \times M}$,

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1(X_1) & \dots & \phi_M(X_1) \\ \vdots & & \vdots \\ \phi_1(X_n) & \dots & \phi_M(X_n) \end{pmatrix},$$

Least squares regression problem:

$$\mathbf{w}_n = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \mathbf{w}, \Phi(X_i) \rangle)^2 = \frac{1}{n} \left\| Y - \Phi \mathbf{w} \right\|^2,$$

with **solution**:

$$\mathbf{w}_n = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{Y},$$

where the matrix $(\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \in \mathbb{R}^{M \times n}$ is the pseudo-inverse of $\boldsymbol{\Phi}$ (see 9 Appendix).

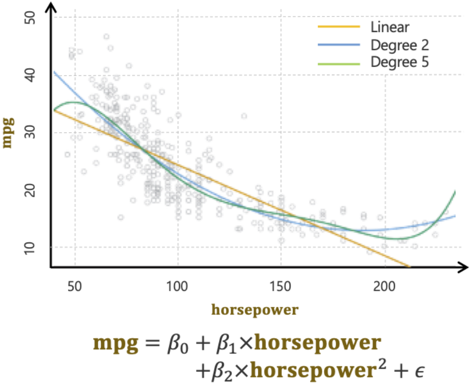
The problem remains linear in w!

Some interesting properties:

- The final function, $f(\mathbf{x}) = \langle \mathbf{w}_n, \Phi(\mathbf{x}) \rangle = \sum_{i=1}^{M} w_{ni}\phi_i(\mathbf{x})$, is linear in the parameter $\mathbf{w}_n$. The functions $\Phi(\mathbf{x})$ allow us to directly incorporate our prior knowledge.
- *Problem:* if we aim to model all polynomials in $\mathbb{R}^d$, then we need d polynomials of degree one (linear functions), $\frac{d(d+1)}{2}$ polynomials of degree two, ... That is, the set of basis functions increases rapidly with d, making this approach unpractical.

Example

Gas mileage dataset: input x is measures the horsepower, output measures the miles per gallon (mpg), $n = 397$.



$$\mathbf{mpg} = \beta_0 + \beta_1 \times \mathbf{horsepower} + \beta_2 \times \mathbf{horsepower}^2 + \epsilon$$

Figure: Figure from EML course (Prof. Vreeken & Valera).

Outline

1 Bibliography

2 Least squares regression

3 Basis functions

4 Bias-variance

5 Gauss-Markov-Theorem

6 Ridge Regression

7 Lasso

8 Summary

9 Appendix

Relation to Bayes optimal function

Solutions $\mathbf{w}_n$ of least squares are estimators for the optimal parameter $\mathbf{w}^*$ (Bayes optimal **linear** function for the squared loss), i.e.,

$$\mathbf{w}^* = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}\big[(Y - \langle \mathbf{w}, X\rangle)^2\big] = \mathbb{E}\big[(Y - \sum_{i=1}^{d} w_i X_i)^2\big],$$

whose optimal solution is given by (note that here X is a d-dimensional column vector random variable):

$$\mathbf{w}^* = \Big(\mathbb{E}[X^T X]\Big)^{-1} \mathbb{E}[X^T Y].$$

The empirical solutions $\mathbf{w}_n$ depend on the training sample $D = (X_i, Y_i)_{i=1}^n$. **Key questions:**

- Is the average estimator $\mathbf{w}_n$ over training samples of size n equal to the optimal $\mathbf{w}^*$?
- How much does the estimator $\mathbf{w}_n$ fluctuate around its average value over all possible training samples of size n?

Bias and Variance of an estimator

Definition

Given a sample $D = (X_i)_{i=1}^n$ and an estimator (also called statistics) $f_n : D \to \mathbb{R}$ of a quantity $f \in \mathbb{R}$. Then, the **bias** of the estimator f_n is defined as

$$\mathrm{Bias}(f_n) = \mathbb{E}_D\left[f_n\right] - f,$$

which corresponds to the difference of the expectation of f_n over all training sets D (all possible i.i.d. training sets of size n) and the true quantity f.

- The estimator f_n is said to be **unbiased** if the bias is zero.
- It is **asymptotically unbiased** if $\lim_{n\to\infty} \mathrm{Bias}(f_n) = 0$.

The **variance** of an f_n is defined as,

$$\mathrm{Var}\, f_n = \mathbb{E}_D\left[(f_n - \mathbb{E}_D[f_n])^2\right],$$

and accounts for the variability of the estimate across different datasets of size n.

Examples for bias and variance

- **The empirical mean** $\mathbb{E}_{\mathrm{P}_n}[X] = \frac{1}{n}\sum_{i=1}^n X_i$ is an **estimator of the true mean** $\mathbb{E}_{\mathrm{P}}[X] = \mathbb{E}[X]$, and it fulfills that:

$$\mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{X_i}\left[X_i\right] = \frac{1}{n} n\mathbb{E}[X] = \mathbb{E}[X] \implies \text{unbiased!}$$

- **The empirical variance** $\mathrm{Var}_{\mathrm{P}_n}[X] = \frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}_{\mathrm{P}_n}[X])^2$ as an estimator of the **true variance** $\mathrm{Var}_{\mathrm{P}}[X] = \mathrm{Var}[X]$, and it fulfills that:

$$\mathbb{E}_D\left[\mathrm{Var}_{\mathrm{P}_n}[X]\right] = \frac{n-1}{n}\mathrm{Var}[X] \implies \text{biased! underestimation!}$$

Instead, the estimator $\frac{1}{n-1}\sum_{i=1}^n (X_i - \mathbb{E}_{\mathrm{P}_n}[X])^2$ for the variance of X is unbiased.

Risk of an estimator

Let's assume a (fixed) estimator f_n of a regression function $f : \mathcal{X} \to \mathcal{Y}$ and the squared loss, so that the Bayes optimal function is $f^*(X) = \mathbb{E}[Y|X]$. Then, the risk (i.e., the expected squared loss) $R(f_n)$ of the estimator f_n is given by:

$$\begin{aligned}
R(f_n) &= \mathbb{E}[(Y - f_n(X))^2] = \mathbb{E}\big[\,\mathbb{E}[(Y - f_n(X))^2|X]\big] \\
&= \mathbb{E}\big[\,\mathbb{E}[(Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - f_n(X))^2|X]\,\big] \\
&= \mathbb{E}\big[\,\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]\,\big] + \mathbb{E}\big[\,\mathbb{E}[(\mathbb{E}[Y|X] - f_n(X))^2|X]\,\big] \\
&\quad + 2\mathbb{E}\big[\,\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - f_n(X))|X]\,\big], \\
&= \mathbb{E}\big[(Y - \mathbb{E}[Y|X])^2\big] + \mathbb{E}\big[(\mathbb{E}[Y|X] - f_n(X))^2\big]
\end{aligned}$$

Interpretation:

- The first term is the **Bayes optimal risk** (often also called noise term), where $f^*(x) = \mathbb{E}[Y|X = x]$ is the Bayes optimal function for the squared loss (note that this term cannot be further reduced).
- The second term measures the **deviation of f_n from the Bayes optimal function**. It is a random quantity since f_n depends on the training data!

Expected risk over all training datasets

Expected risk $\mathbb{E}_D[R(f_n)]$ over all possible training sets D of size n is given by:

$$\mathbb{E}_D\left[R(f_n)\right] = \mathbb{E}\left[(Y - f^*(X))^2\right] + \mathbb{E}_D\left[\mathbb{E}_X[(f^*(X) - f_n(X))^2]\right],$$

Note that the first term is constant and the second (by exchanging the two expectations) can be elaborated as:

$$\begin{aligned}
&\mathbb{E}_D\left[(f_n(X) - f^*(X))^2\right] = \mathbb{E}_D\left[(f_n(X) - \mathbb{E}_D[f_n(X)] + \mathbb{E}_D[f_n(X)] - f^*(X))^2\right] \\
&= \mathbb{E}_D\left[(f_n(X) - \mathbb{E}_D[f_n(X)])^2\right] + \mathbb{E}_D\left[\left(\mathbb{E}_D[f_n(X)] - f^*(X)\right)^2\right] \\
&+ 2\mathbb{E}_D\left[\left(f_n(X) - \mathbb{E}_D f_n(X)\right)\left(\mathbb{E}_D f_n(X) - f^*(X)\right)\right] \\
&= \mathbb{E}_D\left[(f_n(X) - \mathbb{E}_D[f_n(X)])^2\right] + \left(\mathbb{E}_D[f_n(X)] - f^*(X)\right)^2 \\
&= \mathrm{Var}(f_n(X)) + (\mathrm{Bias}(f_n(X)))^2,
\end{aligned}$$

Bias-Variance Decomposition

Finally, the expected risk over all training datasets is thus given by:

$$\mathbb{E}_D\big[R(f_n)\big] = \mathbb{E}_{Y,X}\big[(Y - f^*(X))^2\big] + \mathbb{E}_X\Big[\big(\operatorname{Bias} f_n(X)\big)^2\Big] + \mathbb{E}_X\big[\operatorname{Var} f_n(X)\big],$$

which is written as **(Noise)-Bias-Variance-Decomposition**, i.e.,

$$\boxed{\textbf{expected loss} = \textbf{noise} + \textbf{variance} + \textbf{squared bias.}}$$

Trade-off between **bias** and **variance**
corresponds to
Trade-off between **overfitting** and **underfitting**.

Bias-Variance Decomposition

Finally, the expected risk over all training datasets is thus given by:

$$\mathbb{E}_D\left[R(f_n)\right] = \mathbb{E}_{Y,X}\left[(Y - f^*(X))^2\right] + \mathbb{E}_X\left[\left(\mathrm{Bias}\, f_n(X)\right)^2\right] + \mathbb{E}_X\left[\mathrm{Var}\, f_n(X)\right],$$

which is written as **(Noise)-Bias-Variance-Decomposition**, i.e.,

$$\boxed{\textbf{expected loss} = \textbf{noise} + \textbf{variance} + \textbf{squared bias.}}$$

Trade-off between **bias** and **variance**
corresponds to
Trade-off between **overfitting** and **underfitting**.

For a particular feature vector $\mathbf{x}$, then

- **Noise term:** $\mathbb{E}\left[(Y - f^*(X))^2 | X = \mathbf{x}\right]$,
- **Variance of** f_n: $\mathrm{Var}\, f_n(\mathbf{x}) = \mathbb{E}_D\left[(f_n(\mathbf{x}) - \mathbb{E}_D[f_n(\mathbf{x})])^2\right]$,
- **Bias of** f_n: $\mathrm{Bias}\, f_n(\mathbf{x}) = \mathbb{E}_D[f_n(\mathbf{x})] - f^*(\mathbf{x})$,

where the latter two terms correspond to the variance and the bias of the f_n estimator at $X = \mathbf{x}$.

Bias-Variance of Least Squares

Bias-Variance-Decomposition for the Least-Squares estimator: the bias and variance of $f_n(\mathbf{x}) = \langle \mathbf{w}_n, \mathbf{x} \rangle$ (i.e., the LS prediction of the target value y for a feature vector $\mathbf{x}$) can be written in terms of respectively the bias and the covariance of $\mathbf{w}_n$, i.e.,

$$
\begin{aligned}
\mathrm{Bias}\, f_n(X = \mathbf{x}) &= \mathbb{E}_D[f_n(\mathbf{x})] - f^*(X) = \mathbb{E}_D[\langle \mathbf{w}_n, \mathbf{x} \rangle] - \langle w^*, \mathbf{x} \rangle \\
&= \langle \mathbb{E}_D[\mathbf{w}_n] - w^*, \mathbf{x} \rangle \\
&= \langle \mathrm{Bias}(\mathbf{w}_n), \mathbf{x} \rangle \,, \\
\mathrm{Var}\, f_n(X = \mathbf{x}) &= \mathbb{E}_D\Big[(f_n(\mathbf{x}) - \mathbb{E}_D[f_n(\mathbf{x})])^2 \Big] = \mathbb{E}_D\Big[(\langle \mathbf{w}_n, \mathbf{x} \rangle - \langle \mathbb{E}_D[\mathbf{w}_n], \mathbf{x} \rangle)^2 \Big] \\
&= \mathbb{E}_D\Big[\, \langle \mathbf{w}_n - \mathbb{E}_D[\mathbf{w}_n], \mathbf{x} \rangle^2 \, \Big] \\
&= \mathbb{E}_D\Big[\sum_{i,j=1}^{d} \big(w_{ni} - \mathbb{E}_D[w_{ni}] \big) x_i x_j \big(w_{nj} - \mathbb{E}_D[w_{nj}] \big) \\
&= \mathrm{tr}(\mathbf{x}\mathbf{x}^T \, \mathrm{Cov}(\mathbf{w}_n))
\end{aligned}
$$

Outline

1 Bibliography

2 Least squares regression

3 Basis functions

4 Bias-variance

5 Gauss-Markov-Theorem

6 Ridge Regression

7 Lasso

8 Summary

9 Appendix

Gauss-Markov-Theorem

Theorem (Gauss-Markov theorem)

Suppose that the data obeys the linear model

$$Y = \langle \mathbf{w}, X \rangle + \epsilon,$$

with $\mathbb{E}[\epsilon|X = \mathbf{x}] = 0$, $\mathrm{Var}[\epsilon|X = \mathbf{x}] = \sigma^2$ *and errors at different points are uncorrelated (i.e., noise is homoscedastic). Then,*

- *the* **least squares estimator (LSE)** $\mathbf{w}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ *is* **unbiased**, *and*
- *among all possible unbiased estimators of the LSE* $\mathbf{w}_n$ *has the* **smallest variance**.

Observations

The Gauss-Markov-Theorem is only of limited practical use:

- Model assumption has to be true! In reality linearity assumption is not often fulfilled.
- If the model assumption is correct, then the least squares estimator is the best among all possible **unbiased** estimators.
- However, a slightly biased estimator (e.g. ridge regression or lasso, covered in next lecture) may present a much smaller variance, and thus better expected squared error.
- In other words, biased estimators (with low variance) may be preferred in practice.

Outline

1 Bibliography

2 Least squares regression

3 Basis functions

4 Bias-variance

5 Gauss-Markov-Theorem

6 Ridge Regression

7 Lasso

8 Summary

9 Appendix

Ridge Regression

Ridge regression: Adds to the LSR formulation an L_2-Regularization term.

Definition (Ridge regression)

Given sample $D_n = (X_i, Y_i)_{i=1}^n$, **ridge regression** is formulated as:

$$D_n \mapsto f_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \mathbf{w}, \Phi(X_i) \rangle)^2 + \lambda \|\mathbf{w}\|_2^2.$$

- Provides a regularized version of LSR and thus will be less prone to overfitting.
- Even if the solution of LSR is not unique, Ridge regression has a unique solution.

Solution of ridge regression

For a given dataset $D_n = (\mathbf{x}_i, y_i)_{i=1}^n$, and a basis function Φ si that the feature vectors are transformed and collected in a matrix $\mathbf{\Phi} \in \Re^{n \times d}$. Then, the solution of ridge regression is given by:

$$\mathbf{w}_{n,\lambda} = (\mathbf{\Phi}^T \mathbf{\Phi} + \lambda \mathbb{1}_d)^{-1} \mathbf{\Phi}^T \mathbf{Y}.$$

Properties:

- The solution $\mathbf{w}_{n,\lambda}$ exists and is unique.
- The regularizer $\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2$ corresponds to a Gaussian prior (with zero mean and unit variance) for maximum a posteriori (MAP) estimation, i.e.,
$$p(\mathbf{w}) \propto e^{-\Omega(\mathbf{w})} = e^{-\|\mathbf{w}\|_2^2}.$$

LSR vs. Ridge regression I

The solution of least squared regression:

$$\mathbf{w}_n = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{Y}.$$

The solution of ridge regression:

$$\mathbf{w}_{n,\lambda} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbb{1}_d)^{-1}\boldsymbol{\Phi}^T\mathbf{Y}.$$

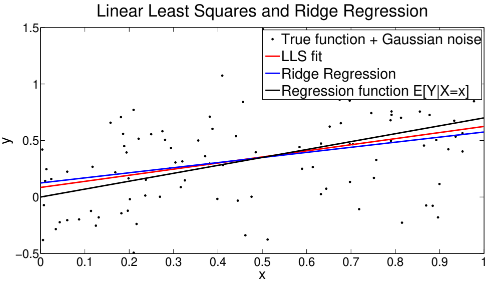


Figure: Linear least squares regression versus linear ridge regression. The regression function is linear. (Image from Prof. Hein)

LSR vs. Ridge regression II

The solution of least squared regression:

$$\mathbf{w}_n = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{Y}.$$

The solution of ridge regression:

$$\mathbf{w}_{n,\lambda} = (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbb{1}_d)^{-1}\mathbf{\Phi}^T\mathbf{Y}.$$

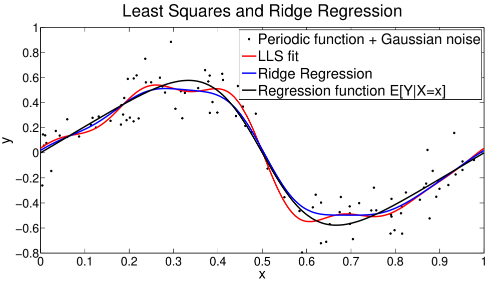


Figure: Comparison of least squares and ridge regression using a set of periodic basis functions. (Image from Prof. Hein)

Geometric interpretation I

Linear LSR uses SVD of X (i.e., $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$), where $\mathrm{rank}(\boldsymbol{\Sigma})$=r,

$$\mathbf{X}\mathbf{w}_n = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V}^T(\boldsymbol{\Sigma}^+)^2\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{Y} = \sum_{i=1}^{r}\mathbf{u}_i\left\langle\mathbf{u}_i, \mathbf{Y}\right\rangle,$$

i.e., the outputs are projected on the basis spanned by $\mathbf{U}$. Above, $\boldsymbol{\Sigma}^+ \in \mathbb{R}^{n\times d}$ is defined as $\boldsymbol{\Sigma}^+_{ij} = \begin{cases} 1/\sigma_i & \text{if } i=j \text{ and } i\leq r, \\ 0 & \text{otherwise} \end{cases}$

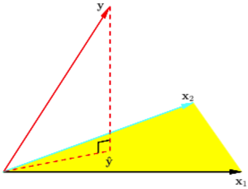


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane spanned by the input vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$. *The projection* $\hat{\mathbf{y}}$ *represents the vector of the least squares predictions*

Figure: Image from ESL.

(Refer to 9 Appendix for further details on SVD.)

Geometric interpretation II

Linear LSR uses SVD of X (i.e., $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$), where $\mathrm{rank}(\boldsymbol{\Sigma})$=r,

$$\mathbf{X}\mathbf{w}_n = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V}^T(\boldsymbol{\Sigma}^+)^2\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{Y} = \sum_{i=1}^{r} \mathbf{u}_i \langle \mathbf{u}_i, \mathbf{Y} \rangle ,$$

i.e., the outputs are projected on the basis spanned by $\mathbf{U}$. Above, $\boldsymbol{\Sigma}^+ \in \mathbb{R}^{n\times d}$ is defined as $\boldsymbol{\Sigma}^+_{ij} = \begin{cases} 1/\sigma_i & \text{if } i = j \text{ and } i \leq r, \\ 0 & \text{otherwise} \end{cases}$

Ridge regression:

$$\mathbf{X}\mathbf{w}_{n,\lambda} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{1}_d)^{-1}\mathbf{X}^T\mathbf{Y} = \sum_{i=1}^{r} \mathbf{u}_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \langle \mathbf{u}_i, \mathbf{Y} \rangle ,$$

i.e., the output $\mathbf{Y}$ is also projected on the basis spanned by $\mathbf{U}$, but here the smaller the singular value σ_i (compared to λ) the larger the **shrinkage** in this direction, or in order words, the smaller is the influence of this direction.

(Refer to 9 Appendix for further details on SVD.)

Outline

1. Bibliography
2. Least squares regression
3. Basis functions
4. Bias-variance
5. Gauss-Markov-Theorem
6. Ridge Regression
7. Lasso
8. Summary
9. Appendix

Lasso

Lasso (least absolute shrinkage and selection operator) corresponds to LSR with L_1-Regularization.

Definition (Lasso)

Given a training sample $D_n = (X_i, Y_i)_{i=1}^n$ with $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$ and the function space $\mathcal{F} = \{\sum_{j=1}^d w_j \phi_j(\mathbf{x}) \,|\, \mathbf{w} \in \mathbb{R}^d\}$ we define **the lasso** as the mapping $\mathcal{A} : D_n \to \mathcal{F}$ with

$$D_n \mapsto \mathbf{w}_n = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (Y_i - \langle \mathbf{w}, \Phi(X_i) \rangle)^2 + \lambda \left\| \mathbf{w} \right\|_1 .$$

Observation: The regularizer $\Omega(\mathbf{w}) = \left\| \mathbf{w} \right\|_1$ corresponds to a Laplace prior for maximum a posteriori (MAP) estimation.

L_1 vs other norms

- A closed form **solution** as for Lasso is generally not available. However, as the objective function to be minimized is convex with respect to $\mathbf{w}$, one may rely on existing and efficient convex optimization techniques to compute the solution to Lasso.
- L_1-norm induces **sparsity** (a lot of elements of $\mathbf{w}_n$ are zero). Sparsity is good: less storage, faster evalution $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$, feature selection.
- Why? L_1-norm is the norm which is "closest" to the "zero norm" ($\|w\|_0 = \sum_{i=1}^{D} \mathbb{1}_{w_i \neq 0}$). The "zero norm" enforces directly sparsity.
- L_2-norm $\|\mathbf{w}\|_2$ penalizes large weights heavily $\Rightarrow$ preference for small weights in all directions (regularizer is **isotropic**). L_1-norm $\|\mathbf{w}\|_1$ penalizes large and small weights "equally" $\Rightarrow$ produces often large weights in few directions.

Comparison: lasso and ridge regression I

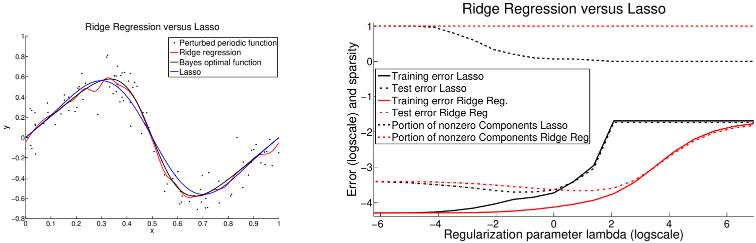


Figure: **Left:** Perturbed training data and regression function in black, we show the solution of ridge regression in blue and of Lasso in red for $\lambda = 1$, **Right:** Behavior of training and test error and number of non-zero components of the weight vector as a function of the regularization parameter λ. (Images from Prof. Hein)

Comparison: lasso and ridge regression II

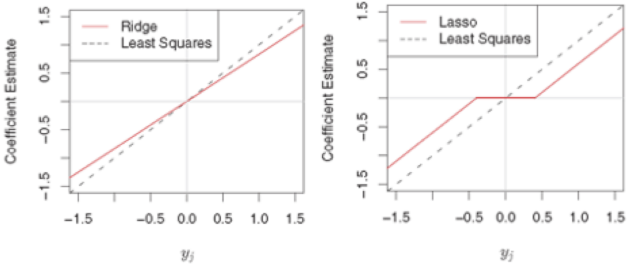


Figure: **Left:** The ridge weights (coefficients) present smaller absolute values than for LSR, but still different that zero. **Right:** In contrast, Lasso triggers a subset of weights directly to zero. (Images from eML)

Comparison: lasso and ridge regression II

Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^\star$. The lasso gives a sparse solution in which $w_1^\star = 0$.

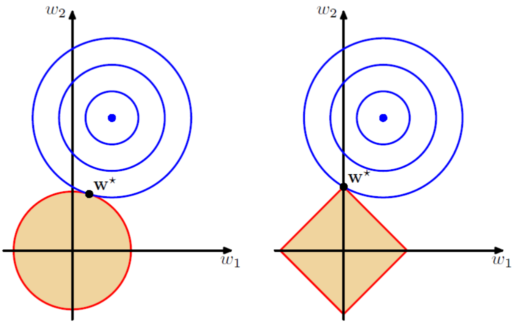


Figure: Image from Bishop.

Regularization functionals

Other regularization functionals: $\Omega(\mathbf{w}) = \sum_{i=1}^{n} |w_i|^p = \|\mathbf{w}\|_p^p$.
$\Rightarrow$ L_2-norm is the only **isotropic** norm in the family of p-norms!

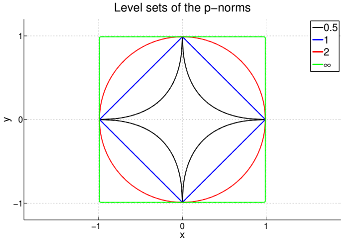


Figure: The level set $\|\mathbf{w}\|_p = 1$ of the p-norms. Note that the $\|\cdot\|_p$ is only a norm for $p \geq 1$, in which case the unit-ball is a convex set. Clearly for $p = 0.5$ the "unit-ball" is not convex.

Outline

1. Bibliography
2. Least squares regression
3. Basis functions
4. Bias-variance
5. Gauss-Markov-Theorem
6. Ridge Regression
7. Lasso
8. Summary
9. Appendix

Summary

Linear regression (and as we will see linear classification too):

- Easy interpretation: feature has a high influence if it has a large weight.
- Linear methods: have possibly high bias but low variance ⇒ can be fit already with only a few training points.
- Often competitive with non-linear methods in high dimensions.
- Using transformations of the input features (**basis functions**) one can easily generate non-linear functions in the input space. Linear methods are *linear* in the parameters, but not necessarily linear in the original input features.

Summary II

Regularized Linear Regression:

- Adding a regularization term, e.g., L2-norm, to LSR formulation allows us to control for overfitting (i.e., for the bias-variance trade-off). It also makes the solution to the regression problem unique, even when the LSR solution is not.
- Lasso induces sparsity in the regression function weighs (i.e., it forces the value of a subset of weights to zero). Thus, Lasso leads to stronger regularization than Ridge regression. However, in contrast to Ridge regression, Lasso does not have close-form solution (efficient convext optimization techniques exist).
- Ridge regression and Lasso can be seen a MAP estimation of the regression weights with, respectively, Gaussian and Laplace priors.

Outline

1 Bibliography

2 Least squares regression

3 Basis functions

4 Bias-variance

5 Gauss-Markov-Theorem

6 Ridge Regression

7 Lasso

8 Summary

9 Appendix

The pseudo-inverse

If $\mathbf{X}$ has rank d, then $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the Moore-Penrose **pseudo inverse** of $\mathbf{X}$.

Definition (Pseudo-inverse)

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then the **pseudo-inverse $\mathbf{A}^+$** of $\mathbf{A}$ is defined as

$$\mathbf{A}^+ = \arg\min_{\mathbf{B} \in \mathbb{R}^{n \times m}} \|\mathbf{AB} - \mathbb{1}_m\|_F^2 ,$$

where $\|\cdot\|_F$ is the **Frobenius norm** ($\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2}$) and $\mathbb{1}_m$ the identity matrix in $\mathbb{R}^m$.

Let $\mathbf{A}$ be an invertible square matrix, then

$$(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{A}^{-1}(\mathbf{A}^T)^{-1}\mathbf{A}^T = \mathbf{A}^{-1}.$$

SVD

The **singular value decomposition** of $\mathbf{A} \in \mathbb{R}^{m \times n}$,

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T,$$

- U is an orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m) \in \mathbb{R}^{m \times m}$, that is $\mathbf{U}^T\mathbf{U} = \mathbb{1}_m$,
- $\mathbf{V}$ is an orthogonal matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathbb{R}^{n \times n}$, that is $\mathbf{V}^T\mathbf{V} = \mathbb{1}_n$,
- $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$ with $\Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i = j \text{ and } i \leq r, \\ 0 & \text{otherwise} \end{cases}.$

  The $\sigma_i > 0$, $i = 1, \ldots, r$ are the **singular values** of $\mathbf{A}$.

The **pseudo inverse** of a matrix A is then given by

$$\mathbf{A}^+ = \mathbf{A}^{-1}(\mathbf{A}^T)^{-1}\mathbf{A}^T = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^T,$$

where $\boldsymbol{\Sigma}^+ \in \mathbb{R}^{n \times m}$ is defined as $\Sigma^+_{ij} = \begin{cases} 1/\sigma_i & \text{if } i = j \text{ and } i \leq r, \\ 0 & \text{otherwise} \end{cases}.$

Proof.

The **pseudo inverse** A^+ is then given by

$$\mathbf{A}^+ = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^T,$$

where $\boldsymbol{\Sigma}^+ \in \mathbb{R}^{n\times m}$ is given by $\Sigma^+_{ij} = \begin{cases} 1/\sigma_i & \text{if } i = j \text{ and } i \le r, \\ 0 & \text{otherwise} \end{cases}$.

Let $\mathbf{A} \in \mathbb{R}^{n\times m}$. Given that $m \le n$ and $rank(\mathbf{A}) = m$, one can write the pseudo inverse $\mathbf{A}^+$ as $\mathbf{A}^+ = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$,

$$\begin{aligned}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T &= (\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)^{-1}\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T = (\mathbf{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\mathbf{V}^T)^{-1}\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T \\ &= \mathbf{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})^{-1}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T = \mathbf{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^T\mathbf{U}^T = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^T.\end{aligned}$$