

CSE 392: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin
Spring 2024

Lecture 3: Least squares regression and kernel methods

Outline

- 1 Least squares regression
- 2 Ridge regression
- 3 Kernel methods

Data fitting - Regression

- We are given,
 - ▶ A data matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with n samples $\{\mathbf{a}_i\}_{i=1}^n \in \mathbb{R}^d$ of d -dimensional features, and
 - ▶ A column vector $\mathbf{b} \in \mathbb{R}^n$ (targets).
- **Data fitting:** Find a functional relation between features and targets wrt. certain loss. General form: For a loss function $\ell(\cdot, \cdot)$, and a function $f(\cdot, \theta)$, where θ are the function parameters over a possible set Θ , we solve

$$\theta^* = \min_{\theta \in \Theta} \sum_{i=1}^n \ell(f(\mathbf{a}_i, \theta), b_i)$$

- **Numerous applications** from scientific computing to machine learning, finance, statistics and many more.

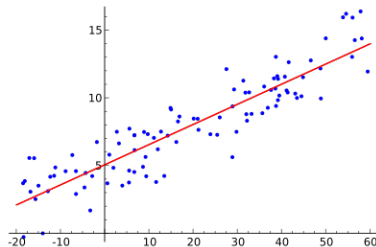
Least squares linear regression

Exercises:

- In the *least-squares* regression problem, assuming $d < n$, we solve:

$$\mathbf{x}^* = \min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2.$$

- A linear function and Euclidean- (ℓ_2) norm (squared) loss function.
- The observed targets, $b_i = \mathbf{a}^\top \mathbf{x} + \varepsilon_i$, for $i = 1, \dots, n$ and ε_i is noise..



Normal equation

The vector \mathbf{x}^* minimizes $\|\mathbf{Ax} - \mathbf{b}\|^2$ if and only if it is the solution of the **normal equations**:

$$\mathbf{A}^\top \mathbf{Ax} = \mathbf{A}^\top \mathbf{b}.$$

Normal equation

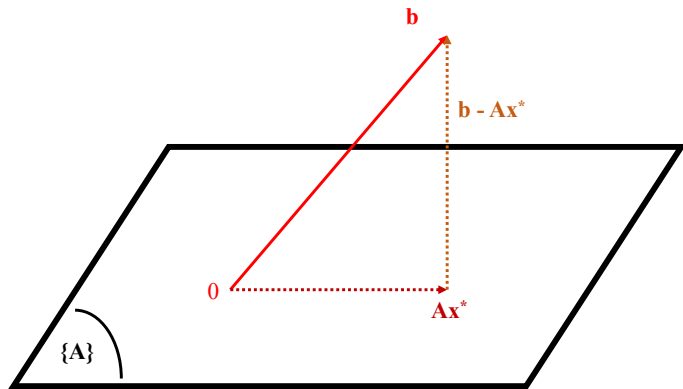
The vector \mathbf{x}^* minimizes $\|\mathbf{Ax} - \mathbf{b}\|^2$ if and only if it is the solution of the **normal equations**:

$$\mathbf{A}^\top \mathbf{Ax} = \mathbf{A}^\top \mathbf{b}.$$

Proof: Consider any $\tilde{\mathbf{x}} = \mathbf{x}^* + \Delta\mathbf{x}$, then we have

$$\begin{aligned}\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|^2 &= \|\mathbf{Ax}^* + \mathbf{A}\Delta\mathbf{x} - \mathbf{b}\|^2 \\ &= \|\mathbf{Ax}^* - \mathbf{b}\|^2 - 2(\mathbf{A}\Delta\mathbf{x})^\top (\mathbf{Ax}^* - \mathbf{b}) + \|\mathbf{A}\Delta\mathbf{x}\|^2 \\ &= \|\mathbf{Ax}^* - \mathbf{b}\|^2 - 2(\Delta\mathbf{x})^\top \underbrace{\mathbf{A}^\top (\mathbf{Ax}^* - \mathbf{b})}_{\nabla_{\mathbf{x}}\ell} + \underbrace{\|\mathbf{A}\Delta\mathbf{x}\|^2}_{\geq 0}\end{aligned}$$

Hence, $\|\mathbf{A}(\mathbf{x}^* + \Delta\mathbf{x}) - \mathbf{b}\|^2 \geq \|\mathbf{Ax}^* - \mathbf{b}\|^2$ for any $\Delta\mathbf{x}$, iff the gradient vector $\nabla_{\mathbf{x}}\ell$ is zero.



\mathbf{x}^* is the best approximation to \mathbf{b} from the subspace $\text{span}\{\mathbf{A}\}$ iff $(\mathbf{b} - \mathbf{Ax})$ is \perp to the whole subspace $\text{span}\{\mathbf{A}\}$. This in turn is equivalent to Normal equations

$$\mathbf{A}^\top (\mathbf{Ax}^* - \mathbf{b}) = 0.$$

Matlab demo

Issue with normal equations

The solution is $\mathbf{x}^* = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b}$.

- **Condition number** of a matrix :

$$\kappa_2(\mathbf{A}) = \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 = \sigma_{\max} / \sigma_{\min}$$

- Then, $\kappa_2(\mathbf{A}^\top \mathbf{A}) = \|\mathbf{A}^\top \mathbf{A}\|_2 \|(\mathbf{A}^\top \mathbf{A})^{-1}\|_2 = (\sigma_{\max} / \sigma_{\min})^2$.

E.g., suppose we have a matrix with spectrum in $[1, \epsilon]$, i.e, $\kappa_2(\mathbf{A}) = 1/\epsilon$.

Then, $\kappa_2(\mathbf{A}^\top \mathbf{A}) = \epsilon^{-2}$.

$\mathbf{A}^\top \mathbf{A}$ could be highly *ill-conditioned*.

Ridge Regression

Ridge Regression or Tikhonov regularization: For a given $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$ the ridge-regression estimator is the minimizer of the problem:

$$\mathbf{x}_{rr} = \arg \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_2^2,$$

where $\lambda > 0$ is a fixed regularization parameter.

The solution is $\mathbf{x}_{rr} = (\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^\top \mathbf{b}$.

We select an appropriate λ such that:

- we have a better conditioned matrix, and
- we avoid *over fitting*.

Bias–variance tradeoff.

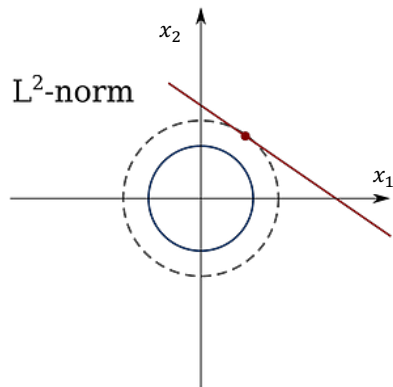
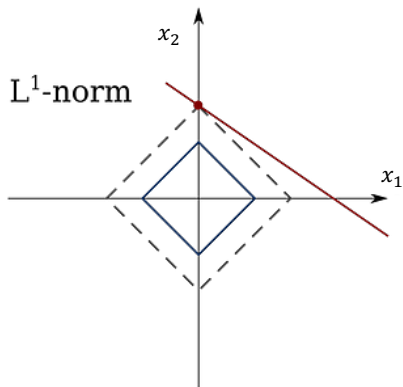
LASSO Regression

Least absolute shrinkage and selection operator, or LASSO , proposed by Tibshirani in 1996, solves the optimization problem:

$$\mathbf{x}_{lasso} = \arg \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1,$$

where $\lambda > 0$ is a fixed regularization parameter.

- The problem is still convex, but is non-smooth.
- Many efficient optimization algorithms have been proposed. E.g., Fast Iterative Shrinkage-Thresholding Algorithm (FISTA), Alternating Direction Method of Multipliers (ADMM).
- Yields a *sparse solution*.



Constraint Regions for LASSO (left) and Ridge Regression (right). Shows why LASSO yields a sparse solution.

Matlab demo II

Feature maps

- Linear regression fits a linear functions to the data.
- However, the functional relation could be “non-linear”.
- **Example:** Consider fitting a cubic function:

$$b = x_3 a^3 + x_2 a^2 + x_1 a + x_0.$$

- We can view the cubic function as a **linear function** over a different set of feature variables. Let the function $\phi : \mathbb{R} \rightarrow \mathbb{R}^4$ be defined as:

$$\phi(a) = [1; a; a^2; a^3].$$

- If $\mathbf{x} = [x_0, x_1, x_2, x_3]$, then

$$b = x_3 a^3 + x_2 a^2 + x_1 a + x_0 = \mathbf{x}^\top \phi(a).$$

- The function ϕ is called the **feature map**.

Kernelization

- Approach to linearize non-linear problems.
- Map rows of \mathbf{A} to $\phi(\mathbf{a}_i)$ in *higher dimension*.
- **Kernel Trick** or kernel substitution: if the input enters an algorithm only in the form of inner products, then we can replace the inner product with some other choice of a kernel.
- **Kernel:** corresponding to the feature map ϕ satisfies:

$$K(\mathbf{a}, \tilde{\mathbf{a}}) = \phi(\mathbf{a})^\top \phi(\tilde{\mathbf{a}})$$

- Kernel is symmetric of its arguments , i.e., $K(\mathbf{a}, \tilde{\mathbf{a}}) = K(\tilde{\mathbf{a}}, \mathbf{a})$.

Kernel properties

Mercer Theorem

Let $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be given. Then for K to be a valid (Mercer) kernel, it is necessary and sufficient that for any $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}, (n < \infty)$, the corresponding kernel matrix is symmetric positive semi-definite.

Kernel properties

Mercer Theorem

Let $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be given. Then for K to be a valid (Mercer) kernel, it is necessary and sufficient that for any $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$, $(n < \infty)$, the corresponding kernel matrix is symmetric positive semi-definite.

Proof: Let the kernel matrix \mathbf{K} be defined as $K_{ij} = \phi(\mathbf{a}_i)^\top \phi(\mathbf{a}_j)$. If K is a valid kernel, then $K_{ij} = \phi(\mathbf{a}_i)^\top \phi(\mathbf{a}_j) = \phi(\mathbf{a}_j)^\top \phi(\mathbf{a}_i) = K_{ji}$, hence symmetric. Also for any vector \mathbf{z} , we have:

$$\begin{aligned} \mathbf{z}^\top \mathbf{K} \mathbf{z} &= \sum_i \sum_j z_i K_{ij} z_j = \sum_i \sum_j z_i \phi(\mathbf{a}_i)^\top \phi(\mathbf{a}_j) z_j \\ &= \sum_i \sum_j z_i \sum_k \phi_k(\mathbf{a}_i) \phi_k(\mathbf{a}_j) z_j = \sum_k \sum_i \sum_j z_i \phi_k(\mathbf{a}_i) \phi_k(\mathbf{a}_j) z_j \\ &= \sum_k \left(\sum_i z_i \phi_k(\mathbf{a}_i) \right)^2 \geq 0. \end{aligned}$$

Kernels as similarity metrics

- Intuitively, when $\phi(\mathbf{a})$ and $\phi(\tilde{\mathbf{a}})$ are close to each other, the kernel $K(\mathbf{a}, \tilde{\mathbf{a}}) = \phi(\mathbf{a})^\top \phi(\tilde{\mathbf{a}})$ should be large.
- Conversely, if they are far apart, $K(\mathbf{a}, \tilde{\mathbf{a}})$ should be small.
- Kernel as a similarity measure of the features.
- **Gaussian Kernel:** Homogeneous kernels defined by the magnitude of distance:

$$K(\mathbf{a}, \tilde{\mathbf{a}}) = \exp\left(-\frac{\|\mathbf{a} - \tilde{\mathbf{a}}\|^2}{2\sigma^2}\right).$$

It corresponds to an infinite dimensional feature map ϕ .

Kernel Ridge Regression

- Kernel methods - do not explicitly define or compute the feature map ϕ . Only compute the kernel function $K(\cdot, \cdot)$.
- In ridge regression, suppose we replace the feature vectors: $\mathbf{a}_i \rightarrow \Phi_i = \phi(\mathbf{a}_i)$ to account for non-linear function relation.
- Now the dimension can be much higher.
- The solution to the ridge regression is, with $\phi(\mathbf{a}_i)$'s as columns of Φ :

$$\mathbf{x}_{kr} = (\Phi\Phi^\top + \lambda\mathbf{I})^{-1}\Phi\mathbf{b} = \Phi (\Phi^\top\Phi + \lambda\mathbf{I})^{-1}\mathbf{b}$$

- Given a new data point \mathbf{a} , the prediction will be:

$$b = \phi(\mathbf{a})^\top \mathbf{x}_{kr} = \mathbf{b}^\top (\Phi^\top\Phi + \lambda\mathbf{I})^{-1} \Phi^\top \phi(\mathbf{a}) = \mathbf{b}^\top (\mathbf{K} + \lambda\mathbf{I})^{-1} \kappa(\mathbf{a}),$$

where $\kappa(\mathbf{a}) = [K(\mathbf{a}_i, \mathbf{a})]_{i=1}^n$.

Questions?