SMAI-S25-09: Regression and Regularization

C. V. Jawahar

IIIT Hyderabad

February 4, 2025

Recap:

- Problems of interest:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
    - (a) Classification (b) Regression
  - Learn Feature Transformations $\mathbf{x}' = \mathbf{W}\mathbf{x}$ or $\mathbf{x}' = f(\mathbf{W}, \mathbf{x})$
- Algorithms/Approaches:
  - Nearest Neighbour Algorithm
  - Linear Classification: $sign(\mathbf{w}^T\mathbf{x})$
  - Decide as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2.
  - Linear Regression
- Performance Metrics:
  - Classification: Accuracy, TP/FP etc., Confusion Matrix; Ranking: Precision, Recall, F-Score, AP
- Supervised Learning:
  - Notion of Training, Validation and Testing
  - Notion of Loss Function, MSE, Regularization (today)
  - Role of Optimization, Convex and non-Convex optimization
  - Closed form solution, Gradient Descent, eigen vector (today)

Linear Regression

Given N examples $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$. Our goal is to find a solution of the form: $y_i = mx_i + c$; or $y_i = w_1 x_i + w_2$ or

$$\hat{y}_i = \mathbf{w}^T \mathbf{x}_i$$

Problem:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

Two Solutions

1. Closed Form Solution

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

2. Gradient Descent Solution

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \eta\nabla J$$

$$\min_{\mathbf{w}} J(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{w}^\mathbf{T}\mathbf{x_i})^2$$

$$\nabla J = \frac{2}{N}\sum_{i=1}^{N}(y_i - \mathbf{w}^\mathbf{T}\mathbf{x_i})(-\mathbf{x_i})$$

Regularization

- Prefer some type of solutions (eg. prefer sparse solutions).
- Frequently used in solving ill-posed problems (eg. avoid singularities such as inverting a near singular matrix)
- A set of methods for reducing overfitting in machine learning models.
- In short, critical for managing model complexity, improving generalization to new data, and addressing specific issues
- Example: Add an additional function to the loss function

$$J = f(\mathbf{w}, \mathbf{x}, y) + \lambda g(\mathbf{w})$$

Regularize with L0, L1 and L2

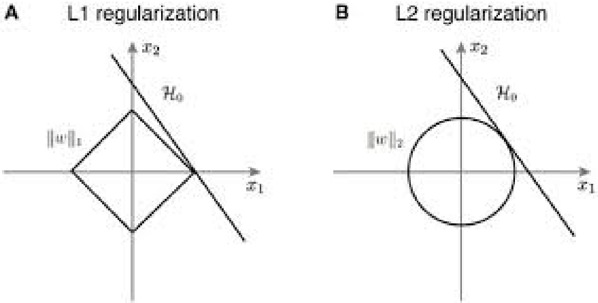
$$J = f(\mathbf{w}) + \lambda||\mathbf{w}||$$

It is common to regularize with $L2$, $L1$ (ideally $L0$; why?) norms.
Recap: L_p norm of of $\mathbf{x}$ is defined as

$$||\mathbf{x}|| = (\sum_{i=1}^{d} |x_i|^p)^{\frac{1}{p}}$$

- L1: (when d=2): $||\mathbf{x}||_1 = |x_1| + |x_2|$
- L2: (when d=2): $||\mathbf{x}||_2 = \sqrt{x_1^2 + x_2^2}$
- The L0 norm of a vector counts the number of non-zero elements in the vector.
- The L-infinity norm, also known as the “max norm,” is a vector norm that measures the maximum absolute value of the vector elements.

Some Intuitive Arguments on Sparsity



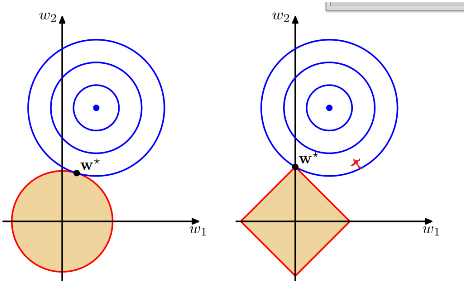
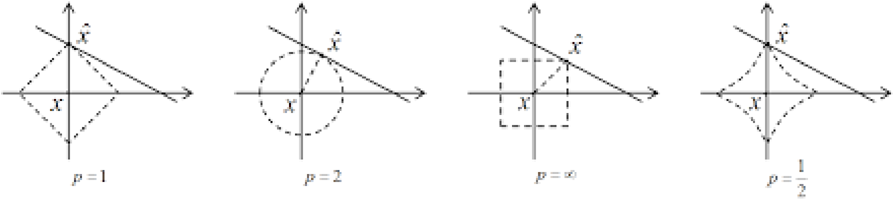
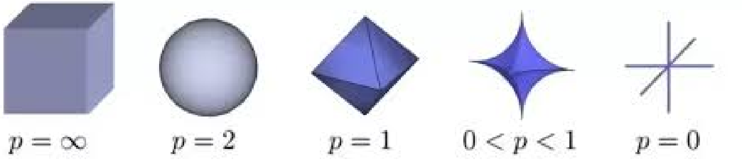


Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^\star$.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

Some Intuitive Arguments on Sparsity





Ridge Regression

In ridge regression, we improve the objective as

$$J = \sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x})^2 + \lambda \sum_{j=1}^{d} w_j^2$$

$$= \sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x})^2 + \lambda||\mathbf{w}||_2^2.$$

Let us come back to ridge regression objective. It can be done as:

$$J(\mathbf{w}) = \frac{1}{N}(\mathbf{Y}^T\mathbf{Y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T(\mathbf{X}^{\mathbf{T}}\mathbf{Y}) + \lambda\mathbf{w}^T\mathbf{w}))$$

Optimization of the above objective function leads to:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{Y}$$

Lasso Regression

Lasso is another variant of the regression where the objective has $L1$ norm of the $\mathbf{w}$ instead of $L2$.

$$
\begin{aligned}
J &= \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x})^2 + \lambda \sum_{i=1}^{d} |w_j| \\
&= \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x})^2 + \lambda ||\mathbf{w}||_1.
\end{aligned}
$$

In Lasso, we not only penalize the high value of some of the coefficients, optimization also leads to zero for irrelevant variables.
The change in the norm of the penalty may seem like only a minor difference, however the behavior of the $L1$-norm is significantly different than that of the $L2$-norm.

Eigen Values and Vectors (recap)

Recap

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$$

Solutions to this:

- Eigen values: $\lambda_1, \lambda_2 \ldots$
- Eigen vector: $\mathbf{v}_1, \mathbf{v}_2 \ldots$

Often λ_is are sorted in decreasing order
Often $\mathbf{v}_i$ are normalized to unit norm.

Optimization problems with Eigen Vectors as Solutions [1]

Problem: Maximize $\mathbf{w}^T\mathbf{A}\mathbf{w}$ such that $\mathbf{w}^T\mathbf{w} = 1$ (or $||\mathbf{w}|| = 1$)

We form an objective with the help of a lagrangian (λ) as

$$J(\mathbf{w}, \lambda) = \mathbf{w}^T\mathbf{A}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1)$$

Differentiating wrt $\mathbf{w}$ and equating to zero leads to:

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{w}$$

Or $\mathbf{w}$ is the eigen vector corresponding to the largest eigen value.

[1] Benyamin Ghojogh et al, Eigenvalue and Generalized Eigenvalue Problems: Tutorial, Arxiv, 2023. An excellent read beyond this course.

Problem 1

We are interested in solving an overdetermined system of homogeneous equations

$$\mathbf{Ax} = \mathbf{0}$$

where $\mathbf{A}$ is $m \times n$.
Problem is formulated as:

$$\arg\min_{\mathbf{x}} ||\mathbf{Ax}|| \quad \text{Subject to: } ||\mathbf{x}|| = 1$$

Solution to this problem is:

- (a) Eigen Vector corresponding to the largerst eigen value of $\mathbf{AA}^T$.
- (b) Eigen Vector corresponding to the largerst eigen value of $\mathbf{A}^T\mathbf{A}$.
- (c) Eigen Vector corresponding to the smallest eigen value of $\mathbf{AA}^T$.
- (d) Eigen Vector corresponding to the smallest eigen value of $\mathbf{A}^T\mathbf{A}$.

Read later[2]

[2] https://cw.fel.cvut.cz/old/_media/courses/a3m33iro/overdetermined_homogenous_linear_equations.pdf

Problem 2

- Consider the $d \times N$ data matrix $\mathbf{X}$ with every column as data elements. Show that $\mathbf{A} = \mathbf{X}\mathbf{X}^T$ is Symmetric. Show that $\mathbf{A}$ is PSD. [3]
- Consider a data matrix $\mathbf{X}$ where every column is mean subtracted samples. We compute $\mathbf{A} = \mathbf{X}\mathbf{X}^{\mathbf{T}}$. Then $\mathbf{A}$ (agree/disagree)
  - is a $d \times d$ matrix
  - is a $N \times N$ matrix
  - is a symmetric matrix
  - is a scaled version of the covariance matrix

[3] A matrix A is PSD if $z^T A z \geq 0 \;\; \forall z \in R^d$.

Problem 3

Consider a square matrix $\mathbf{A}$ with eigen values λ_i and eigen vectors $\mathbf{v}_i$. Then for $\mathbf{A}^T$,

- (a) Eigen values and eigen vectors are the same as that of $\mathbf{A}$.
- (b) Eigen values are the same. Eigen vectors are $\mathbf{v}^T$.
- (c) Eigen values are $\frac{1}{\lambda_i}$
- (d) We can comment for symmetric matrix $\mathbf{A}$. But not for other square matrices.
- (e) None of the above.

The the eigen values of $\mathbf{A}$ are λ_i, what are the eigen values of $\alpha\mathbf{A}$, where α is a scalet.

- (a) λ_i itself.
- (b) $\frac{\alpha}{\lambda_i}$
- (c) $\alpha\lambda_i$
- (d) $\frac{1}{\alpha}\lambda_i$
- (e) None of the above.

Problem 4: Probabilistic Interpretation of LR

We know that the problem is modelled as

$$y_i = \mathbf{w}^T\mathbf{x} + \epsilon_i$$

We assume that the ϵ_i are distributed IID (Independently and Identically Distributed) according to a Gaussian/Normal distribution with zero mean and variance σ^2

$$p(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$$

This leads to

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \mathbf{w}^T\mathbf{x})^2}{2\sigma^2}\right)$$

This represents the distribution of y_i given $\mathbf{x_i}$ and parameterized by $\mathbf{w}$. We model this as $\mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2)$

Problem 4: Probabilistic Interpretation of LR

Our problem is now to find the MLE estimate of $\mathbf{w}$ from the data. Given the data $(\mathbf{X}, Y)$, we can do a maximum likelihood estimate of $\mathbf{w}$.

$$L(\mathbf{w}) = L(\mathbf{w}, \mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \mathbf{w}^T\mathbf{x})^2}{2\sigma^2}\right)$$

Instead of max $L(\mathbf{w})$, we maximize the log-likelihood. (why? correct?)

$$l(\mathbf{w}) = \log\ L(\mathbf{w}) = \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \mathbf{w}^T\mathbf{x})^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \mathbf{w}^T\mathbf{x})^2}{2\sigma^2}\right) = K\frac{1}{2}\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x})^2$$

Where K is a constant. (why? expression?) This imply now that the MSE objective is only a scaled version of the MLE estimate objective!!.

$$\Rightarrow \min_{\mathbf{w}} J(\mathbf{w}) = \min_{\mathbf{w}} \sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$$