04 - Linear Regression

Ethan Graham

20th June 2023

Reminder: Supervised Classification

We want to minimize: $E(\mathbf{w}) = \sum_{n=1}^{N} L(y(\mathbf{x_n}; \mathbf{w}), t_n)$ Where we have: - x: feature vector - w: model parameters - t: label - y: predictor - L loss function - E error function

ML is an optimization problem!

Line parametrisation

Mathematically we express a $(2\,D)$ line as $y(x; \mathbf{w}) = w_0 + w_1 \cdot x$ Given N pairs $\{x_i, t_i\}$, we want to find the line that most closely fits the observations. Essentially, assuming D dimensional space, we are looking for optimal line parameters $w_0, ..., w_D$ Once again, the natural measure of distance is Euclidean distance. In practice we mostly use the **squared euclidean distance** however. This penalizes greater distances more harshly. This also allows us to express the problem as a **least-squares** problem *recall Nummet*

Once we have found optimal parameters $\mathbf{w} = w\star_0, ...w\star_D$, we predict with the formula:

$$y_t = w_0 + \mathbf{w_{1 \to D}} \cdot \mathbf{x_t}$$

Or

$$y_t = \mathbf{w}^T \begin{bmatrix} 1 \\ x_0 \\ ... \\ x_D \end{bmatrix}$$

In D-dimensional space, we no longer try to fit a line but a **hyperplane** to our dataset.

Because the output is still one dimensional, we can use the least-squares formulation from previously

$$\underset{w}{min} \frac{1}{N} \sum_{i=1}^{N} (\mathbf{w^T} \cdot \mathbf{x_i} - t_i)^2$$

Ridge Regression

In ridge regression, we add a penalty term to the cost function that is equal to a coefficient λ times the magnitude of the term. This penalizes terms with large magnitudes, where λ determines how much we penalize the terms. This helps with overfitting and avoiding multicollinearity.

let M be the size of the dataset, and p the dimension of the feature vectors. Then the cost function becomes:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

Where w_i, $i = 1, ..., p$ are regression coefficients.

In a closed form least-squares solution, we add a regularization term or L_2 regularization term. This prevents the regression coefficients from becoming too large. We get the following ordinary least-squares function:

$$\mathrm{m}inimize : \left\|\mathbf{y} - X\beta\right\|^2 + \lambda \left\|\beta\right\|^2$$

Where $\mathbf{y}$ represents the vector of observed responses, X the matrix of predictor variables, β the vector of regression coefficients ($\mathbf{w}$ previously) and λ the regularization parameter which *shrinks* the parameters. The larger λ is, the more *shrinkage*.

A few short notes on optimization

There are a few methods that we commonly use for optimizing a minimization problem such as this

- Gradient Descent *(applicable to many problems)*
- Closed Form Solution *(only really applicable to linear regression)*

Linear Regression Closed-Form Solution

We want to find a solution $\nabla_w E(w) = 0$ impliying that have found a minimum. Expansion yields the following formula for finding the optimal line parameters $w\star$ $w\star = (\mathbf{X^T X})^{-1}\mathbf{X^T t}$. This is just the lear-squares solution $\mathbf{X^T t} = \mathbf{X^T X w}\star$

Evaluation Metrics for Regression

One of the common evaluation metrics is the **mean-squared error**

$MSE = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - t_i)^2$

Another common metric is the **root mean-squared error** where we simply take the square root of MSE.

We also use **mean absolute error**

$MAE = \frac{1}{N_t} \sum_{i=1}^{N_t} |y_i - t_i|$

or **mean absolute percentage error**

$MAPE = \frac{1}{N_t} \sum_{i=1}^{N_t} \left| \frac{y_i - t_i}{t_i} \right|$

Taking percentage with respect to the true value may be easier to interpret.

Final notes on interpreting a linear model

It is important to be careful of different magnitudes in the features since it may lead to incorrect predictions *(some features overpower others).* We can address this by normalizing the dataset.