From the CIS520 Machine Learning

Lectures: Regression

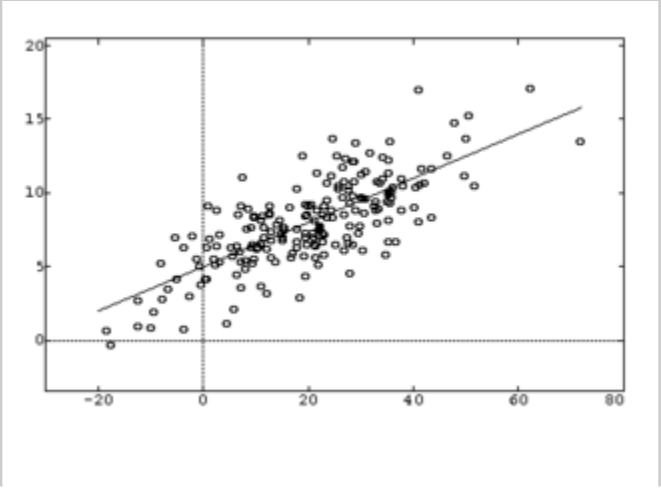
On this page... (hide)

Continuous variable prediction

Suppose now instead of just measuring a single variable of interest, like the amount of coffee in your cup you buy each morning at your coffeeshop, we have some other possibly relevant measurements that can help us predict it. For example, how busy the place is, how good is the mood of the barista, the price of gas. As many smart people reportedly have said, “Prediction is hard, especially about the future.” But we’re going to try anyway, in the simplest way possible, using linear regression.

Linear regression

Consider first the case of a single variable of interest y and a single predictor variable x. The predictor variables are called by many names: covariates, inputs, features; the predicted variable is often called response, output, outcome.



A regression dataset in one dimension

We have some data $D = \{x_i, y_i\}$ and we assume a simple linear model of this data with Gaussian noise:

$$Y = aX + b + Z \quad \text{with} \quad Z \sim \mathcal{N}(0, \sigma^2)$$

Now, $Y_i \sim \mathcal{N}(aX_i + b, \sigma^2)$, so let's do MLE for a and b.

$$\log P(D|a,b) = -n \log(\sigma\sqrt{2\pi}) - \sum_i \frac{(y_i - ax_i - b)^2}{2\sigma^2}$$

Note that the negative log-likelihood is simply residual squared error $\frac{1}{2}\sum_i (y_i - ax_i - b)^2$ (divided by σ^2). Setting derivatives wrt to a and b to 0, we get linear equations (called normal equations) in a,b:

$$\sum_i \frac{x_i(y_i - ax_i - b)}{\sigma^2} = 0 \rightarrow a\sum_i x_i^2 + b\sum_i x_i = \sum_i y_i x_i$$

$$\sum_i \frac{y_i - ax_i - b}{\sigma^2} = 0 \rightarrow a\sum_i x_i + bn = \sum_i y_i$$

For this problem, we can solve for a and b by hand (please do, it's a good exercise), but in general we will have many predictors (inputs).

Multivariate case

Now our features are vectors of dimension m which we will denote in boldface as $\mathbf{x}_i$ and we will index the j-th feature of i-th example as x_{ij}. So we will arrange all the examples in a matrix $\mathbf{X}$, with rows $\mathbf{x}_i$, as follows:

$$\begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix}$$

Similarly, we arrange all the outcomes in a column vector $\mathbf{y}$ of height n. We assume a linear model with m coefficients, which we arrange in a column vector $\mathbf{w}$:

$$y_i = \sum_j w_j x_{ij} + \mathcal{N}(0, \sigma^2) = \mathbf{x}_i \mathbf{w} + \mathcal{N}(0, \sigma^2)$$

As above, maximizing log likelihood is equivalent (up to a scaling and additive term) to minimizing residual squared error:

$$\log P(D \mid \mathbf{w}, \sigma) = -n \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_i (y_i - \mathbf{x}_i \mathbf{w})^2$$

$$= -n \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w})$$

r = y - y_hat

Before computing the MLE, let's define the vector $\mathbf{r} = (\mathbf{y} - \mathbf{X}\mathbf{w})$, as the residual or error vector. The layout of all these vectors and matrices together is:

$$\begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$$

What happened to the constant term?

In the single feature case, we included b, the constant term. We dropped it for the multivariate case. Fortunately, it's easy to add it back by simply including an extra feature that is always 1.

What about polynomial or non-linear regression?

Suppose that dependence on some original feature (barista's mood, gas price) is not linear, but quadratic or cubic or sinusoidal or other non-linear function, $f(x)$. Simple: we just add $f(x)$ as a new feature.

MLE

X T X is a p x p matrix. if p > n, there are more predictors than there are observations

Taking (vector) derivatives of the residual squared error with respect to the parameters, we get: (if you need a refresher and a good quick reference on matrix manipulations, the *Matrix Cookbook* [1] is very handy)

$$\frac{\partial}{\partial \mathbf{w}} \frac{\mathbf{r}^\top \mathbf{r}}{2} = -\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) = 0 \quad \rightarrow \quad \mathbf{w}_{MLE} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Of course, we need $\mathbf{X}^\top \mathbf{X}$ to be invertible. When is that a problem? When the features are not linearly independent. *i.e. price in one column and tax in the other*

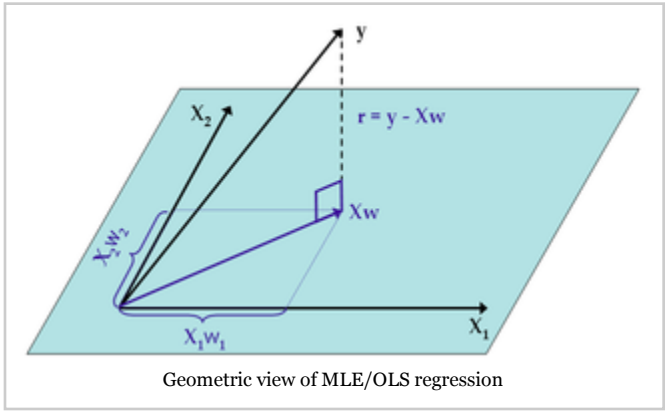
if you have fewer observations than predictors, the matrix is singular and we cannot take the inverse

The regression coefficients we derived from MLE are often called Ordinary Least Squares (OLS). There are many other variants, and we'll see some soon. If you're wondering what the MLE/OLS is doing geometrically, there is a nice interpretation: If we view the outcomes $\mathbf{y}$ as a vector in Euclidean space, and similarly view each input variable (i.e. each column of the data matrix $\mathbf{X}$) as a vector the same space, then the least squares regression is equivalent to an *orthogonal projection* [2] of $\mathbf{y}$ onto the subspace spanned by the input variables (or equivalently, to the column space of X). The projected vector is the prediction $\mathbf{X}\mathbf{w}_{MLE} = \mathbf{P}\mathbf{y}$ where $\mathbf{P} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is the projection matrix. This relationship is depicted for three examples and two input

This means the MLE often gives you a divide-by-zero error if X T X is singular

one thing you can do is "feature selection" whereby you remove predictors by choosing one of two that are strongly correlated

variables (in the picture $\mathbf{w}$ refers to $\mathbf{w}_{MLE}$):



Geometric view of MLE/OLS regression

MAP

And now for the Bayesian way, of course. We will assume a prior on the parameters we are trying to estimate, in this case, $\mathbf{w}$. We will pick the simplest and most convenient one, the Gaussian. Since we usually don't really know much a priori about $\mathbf{w}$, we will use a prior mean of 0 and a standard deviation λ for each parameter:

j = {1…p}

$$w_j \sim \mathcal{N}(0, \lambda^2) \;\text{ so }\; P(\mathbf{w}) = \prod_j \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-w_j^2}{2\lambda^2}}$$

normalization constant

Where do we get λ? We use **cross-validation**. So now to get MAP we need:

$$\begin{aligned}\arg\max_{\mathbf{w}} \log P(\mathbf{w} \mid D, \sigma, \lambda) &= \arg\max_{\mathbf{w}} \left(\log P(D \mid \mathbf{w}, \sigma) + \log P(\mathbf{w} \mid \lambda)\right) \\ &= \arg\min_{\mathbf{w}} \left(\frac{1}{2\sigma^2}\mathbf{r}^\top\mathbf{r} + \frac{1}{2\lambda^2}\mathbf{w}^\top\mathbf{w}\right)\end{aligned}$$

error *L2 norm squared*

Note the effect of the prior: it penalizes large values of $\mathbf{w}$.

Taking derivatives, we get:

$$\frac{\partial}{\partial \mathbf{w}}\left(\frac{\mathbf{r}^\top\mathbf{r}}{2\sigma^2} + \frac{\mathbf{w}^\top\mathbf{w}}{2\lambda^2}\right) = -\frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{1}{\lambda^2}\mathbf{w} = 0$$

which leads to:

covariance matrix

$$\mathbf{w}_{MAP} = (\mathbf{X}^\top\mathbf{X} + \frac{\sigma^2}{\lambda^2}\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

take a symmetric matrix, add something on the diagonal, and it's now invertible!
this is multivariate ridge regression, or "shrinkage" (as a general term)

a Gaussian prior is a very strong assumption that no values are very extreme

There are several trends to observe. As $\lambda \to \infty$, which means our prior is getting broader, MAP reduces to MLE. As the number of samples increases while λ and σ stay

An L2 norm shrinks "big" things (things with a large squared value) a lot, and shrinks small things by much less
*argmin_w [Err + \lambda * w^2]*

fixed, the effect of the prior is vanishes, since the entries of $\mathbf{X}^\top \mathbf{X}$ grow linearly with n. As the noise parameter σ decreases, the effect of the prior also vanishes. As the noise parameter σ increases, MAP shrinks towards prior mean of 0, $\mathbf{w}_{MAP} \to 0$.

The MAP regression above (assuming a zero-mean Gaussian prior) is often called Ridge Regression, where ridge refers to the quadratic penalty of the weights. There are many other priors you can assume, of course, and we'll also see some soon.

Links

1. **www.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf**
2. **en.wikipedia.org/wiki/Projection_%28linear_algebra%29#Orthogonal_projections**