

Machine Learning

Textbook: Machine Learning Refined, by Watt, Borhani, and Katsaggelos, 2nd edition:
MLRefined.com

Previous Lecture

- ◆ SSE, MSE, RMSE, and MAE summarize prediction error in different ways.
- ◆ RMSE and MAE have the same units as the response variable; MSE does not.
- ◆ $R^2 = 1 - \text{SSE}/\text{SST}$ compares the fitted model with the mean-response baseline.
- ◆ A high R^2 does not guarantee that a model is appropriate or predictive.
- ◆ Residual plots can reveal curvature, changing spread, and unusual observations hidden by summary metrics.
- ◆ Training error measures fit to seen data; predictive performance must ultimately be assessed on unseen data.



§5.5, 5.9 - Weighted and Probabilistic Regression

So far:

every observation has contributed equally to the regression loss.

But observations need not be equally reliable.

Today:

weighted least squares,
and a probabilistic explanation for squared error.

Why Give Different Observations Different Weights?

Suppose two measurements have different precision.

One comes from a very accurate instrument.

Another has much greater measurement uncertainty.

Should a residual of size 2
count equally for both?

Not necessarily.

Introduce a positive weight

$$\beta_p > 0$$

for observation p .

Large β_p :

observation matters more in the fit.

Small β_p :

observation matters less.

Weighted Least Squares

Ordinary least squares:

$$g(\mathbf{w}) = \frac{1}{P} \sum_{p=1}^P r_p^2.$$

Weighted least squares:

$$g_W(\mathbf{w}) = \frac{1}{P} \sum_{p=1}^P \beta_p r_p^2.$$

where

$$r_p = f(\mathbf{x}_p; \mathbf{w}) - y_p.$$

Multiplying every weight by the same positive constant
does not change the minimizing parameter vector.

Only relative weights matter.

A Constant Model First

Suppose

$$\hat{y} = w.$$

Weighted squared loss:

$$g_W(w) = \sum_{p=1}^P \beta_p (w - y_p)^2.$$

Differentiate:

$$g'_W(w) = 2 \sum_{p=1}^P \beta_p (w - y_p).$$

At the minimum:

$$\sum_{p=1}^P \beta_p (w - y_p) = 0.$$

Therefore

$$w^* = \frac{\sum_{p=1}^P \beta_p y_p}{\sum_{p=1}^P \beta_p}.$$

Weighted least squares chooses a **weighted mean**.

A Measurement-Precision Example

Suppose three measurements are

$$10, 11, 20.$$

Their standard deviations are known to be

$$\sigma_1 = 1, \quad \sigma_2 = 1, \quad \sigma_3 = 3.$$

Use weights

$$\beta_p = \frac{1}{\sigma_p^2}.$$

Thus

$$\beta_1 = 1, \quad \beta_2 = 1, \quad \beta_3 = \frac{1}{9}.$$

The third measurement is much less precise.

Weighted Mean Versus Ordinary Mean

Ordinary mean:

$$\frac{10+11+20}{3} = \frac{41}{3} \approx 13.67.$$

Weighted mean:

$$\begin{aligned} w^* &= \frac{10+11+\frac{1}{9}(20)}{1+1+\frac{1}{9}} \\ &= \frac{209/9}{19/9} \\ &= 11. \end{aligned}$$

The uncertain observation at 20
has much less influence.

The weighting rule has changed the fitted model.

Weighted Least Squares in Matrix Form

Recall

$$\mathbf{r} = X\mathbf{w} - \mathbf{y}.$$

Create the diagonal weight matrix

$$W = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_P \end{bmatrix}.$$

Then

$$g_W(\mathbf{w}) = \frac{1}{P} (X\mathbf{w} - \mathbf{y})^T W (X\mathbf{w} - \mathbf{y}).$$

This is the weighted version of

$$\|X\mathbf{w} - \mathbf{y}\|^2.$$

Differentiate the Weighted Cost

Gradient:

$$\nabla g_W(\mathbf{w}) = \frac{2}{P} X^T W (X\mathbf{w} - \mathbf{y}).$$

Set equal to zero:

$$X^T W (X\mathbf{w} - \mathbf{y}) = \mathbf{0}.$$

Therefore

$$X^T W X \mathbf{w} = X^T W \mathbf{y}.$$

These are the **weighted normal equations**.

Ordinary least squares is the special case
 $W = I$.

A Weighted Regression Line

Consider

x	0	1	2
y	1	2	6
β	1	1	1/4

The last observation receives less weight.

Design matrix:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}.$$

Weight matrix:

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}.$$

Build the Weighted Normal Equations

Compute

$$X^T W X = \begin{bmatrix} 9/4 & 3/2 \\ 3/2 & 2 \end{bmatrix}.$$

Also

$$X^T W \mathbf{y} = \begin{bmatrix} 9/2 \\ 5 \end{bmatrix}.$$

Thus

$$\begin{bmatrix} 9/4 & 3/2 \\ 3/2 & 2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 9/2 \\ 5 \end{bmatrix}.$$

Solution:

$$w_0 = \frac{2}{3}, \quad w_1 = 2.$$

Weighted line:

$$\hat{y} = \frac{2}{3} + 2x.$$

Compare With Ordinary Least Squares

If all three points receive equal weight,

ordinary least squares gives

$$\hat{y} = 0.5 + 2.5x.$$

Weighted least squares gave

$$\hat{y} = \frac{2}{3} + 2x.$$

The high value

(2, 6)

pulls the ordinary least-squares slope upward more strongly.

Reducing its weight reduces that influence.

Weighted Least Squares Is Still Ordinary Least Squares

Let

$$W^{1/2} = \text{diag}(\sqrt{\beta_1}, \dots, \sqrt{\beta_P}).$$

Then

$$W = (W^{1/2})^T W^{1/2}.$$

So

$$\begin{aligned} & (X\mathbf{w} - \mathbf{y})^T W (X\mathbf{w} - \mathbf{y}) \\ &= \|W^{1/2}X\mathbf{w} - W^{1/2}\mathbf{y}\|^2. \end{aligned}$$

Define

$$\begin{aligned} \tilde{X} &= W^{1/2}X, \\ \tilde{\mathbf{y}} &= W^{1/2}\mathbf{y}. \end{aligned}$$

Then weighted least squares becomes

$$\min_{\mathbf{w}} \|\tilde{X}\mathbf{w} - \tilde{\mathbf{y}}\|^2.$$

Same least-squares geometry after rescaling the rows.

Where Should the Weights Come From?

Weights should represent real information.

Examples:

different measurement precisions;

averages based on different sample sizes;

known differences in noise variance.

Weights should **not** be chosen merely

to force the fitted line toward a preferred answer.

A common statistical choice:

$$\beta_p = \frac{1}{\sigma_p^2},$$

where σ_p^2 is the variance of observation p .

Why this particular formula?

Probability gives the answer.

A Probabilistic Regression Model

Instead of saying only

$$y = f(\mathbf{x}; \mathbf{w}),$$

suppose

$$y_p = f(\mathbf{x}_p; \mathbf{w}) + \varepsilon_p.$$

The model gives the systematic part.

ε_p represents unpredictable noise.

Assume first:

$$\varepsilon_p \sim N(0, \sigma^2),$$

independently for each observation.

Same variance for every observation.

Conditional Distribution of the Response

If

$$y_p = f(\mathbf{x}_p; \mathbf{w}) + \varepsilon_p$$

and

$$\varepsilon_p \sim N(0, \sigma^2),$$

then

$$y_p \mid \mathbf{x}_p \sim N(f(\mathbf{x}_p; \mathbf{w}), \sigma^2).$$

The model prediction

$$f(\mathbf{x}_p; \mathbf{w})$$

is the center of a probability distribution.

Regression is now giving more than a point prediction.

It is describing uncertainty around the prediction.

The Gaussian Density

For one observation:

$$\begin{aligned} p(y_p \mid \mathbf{x}_p, \mathbf{w}, \sigma) \\ = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_p - f(\mathbf{x}_p; \mathbf{w}))^2}{2\sigma^2}\right]. \end{aligned}$$

A small residual:

larger density.

A large residual:
smaller density.

The squared residual appears automatically.

Likelihood

We observed the data

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_P, y_P).$$

For fixed observed data,
view the joint density as a function of the parameters.

Assuming independent errors:

$$L(\mathbf{w}) = \prod_{p=1}^P p(y_p \mid \mathbf{x}_p, \mathbf{w}, \sigma).$$

This function is the **likelihood**.

Maximum likelihood chooses parameters
that make the observed data most compatible with the model.

Products Are Awkward

Take logarithms.

Because log is increasing:
maximizing L
is equivalent to maximizing
 $\log L$.

For equal Gaussian variance:

$$\begin{aligned} \log L(\mathbf{w}) \\ = -\frac{P}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{p=1}^P [y_p - f(\mathbf{x}_p; \mathbf{w})]^2. \end{aligned}$$

The first term does not depend on $\mathbf{w}$.

Maximum Likelihood Becomes Least Squares

To maximize
 $\log L(\mathbf{w})$,

we must minimize

$$\sum_{p=1}^P [y_p - f(\mathbf{x}_p; \mathbf{w})]^2.$$

Therefore

Gaussian noise + maximum likelihood $\Rightarrow$ least squares.

Squared error is not merely computationally convenient.

It has a probabilistic interpretation.

Unequal Variances

Now suppose

$$\varepsilon_p \sim N(0, \sigma_p^2).$$

The observations have different noise levels.

Then

$\log L(\mathbf{w})$
contains

$$-\frac{1}{2} \sum_{p=1}^P \frac{[y_p - f(\mathbf{x}_p; \mathbf{w})]^2}{\sigma_p^2}$$

plus terms that do not depend on $\mathbf{w}$
when the σ_p are known.

So maximizing likelihood means minimizing

$$\sum_{p=1}^P \frac{r_p^2}{\sigma_p^2}.$$

Therefore

$$\beta_p = \frac{1}{\sigma_p^2}.$$

Weighted least squares follows directly from unequal Gaussian uncertainty.

Standard Deviation Versus Weight

Because

$$\beta_p = \frac{1}{\sigma_p^2},$$

if standard deviation doubles:

$$\sigma \rightarrow 2\sigma,$$

then weight becomes

$$\frac{1}{(2\sigma)^2} = \frac{1}{4} \frac{1}{\sigma^2}.$$

Twice the noise standard deviation

$\Rightarrow$

one quarter the weight.

Precision is inverse variance.

Homoscedastic and Heteroscedastic

Equal variance:

$$\sigma_p^2 = \sigma^2$$

for every observation.

This is called **homoscedasticity**.

Different variances:

σ_p^2
changes across observations.

This is called **heteroscedasticity**.

The fan-shaped residual pattern from Lesson 1.12
can be evidence of heteroscedasticity.

Weighted regression is one possible response
when the variance structure can be modeled or estimated.

Estimating a Common Noise Level

Suppose the Gaussian model uses one unknown σ^2 .

After fitting $\mathbf{w}$ by least squares,

the maximum-likelihood estimate is

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{\text{SSE}}{P}.$$

Thus

$$\hat{\sigma}_{\text{MLE}} = \sqrt{\text{MSE}}.$$

That is exactly the training RMSE
when MSE uses denominator P .

A classical unbiased variance estimate uses a degrees-of-freedom correction.

But the maximum-likelihood denominator is P .

Point Prediction Versus Distribution

Ordinary regression statement:

$$\hat{y} = f(\mathbf{x}; \mathbf{w}).$$

Probabilistic statement:

$$Y \mid \mathbf{x} \sim N(f(\mathbf{x}; \mathbf{w}), \sigma^2).$$

The first gives a predicted center.

The second also describes uncertainty around that center.

This idea will become especially important
when outputs are categories rather than numbers.

The Noise Model Is an Assumption

Least squares corresponds naturally to Gaussian noise.

But not every application has Gaussian errors.

Large outliers:

Gaussian squared-error modeling may be too sensitive.

Asymmetric errors:

a symmetric Gaussian model may be inappropriate.

Changing variance:

one common σ^2 may be inappropriate.

Choosing a loss function

can be viewed as choosing assumptions
about how prediction errors behave.

Weighted Least Squares in Python

Use the weighted line example:

```
import numpy as np

x = np.array([0., 1., 2.])
y = np.array([1., 2., 6.])

X = np.column_stack((np.ones(len(x)), x))

weights = np.array([1., 1., 0.25])
W = np.diag(weights)
```

New Python idea:

`np`

creates a diagonal matrix.

Solve:

```
A = X.T @ W @ X
b = X.T @ W @ y

w = np.linalg.solve(A, b)
print(w)
```

Output:

```
[0.66666667 2.          ]
```

Solve by Rescaling Instead

For better numerical practice,

transform the rows and use least squares directly.

```
sqrt_weights = np.sqrt(weights)

X_scaled = X * sqrt_weights.reshape(-1, 1)
y_scaled = y * sqrt_weights

w, residuals, rank, singular_values = \
    np.linalg.lstsq(X_scaled, y_scaled, rcond=None)
```

Again:

```
[0.66666667 2. ... ]
```

New Python idea:

```
reshape
```

turns a one-dimensional array into a column
so each row of X receives the correct scale factor.

Gaussian Negative Log-Likelihood in Python

Suppose each observation has known standard deviation:

```
sigma = np.array([1., 1., 2.])
```

Define

```
def neg_log_likelihood(w):  
    residuals = X @ w - y  
  
    return 0.5*np.sum(  
        np.log(2*np.pi*sigma**2)  
        + residuals**2/sigma**2  
    )
```

The terms involving
 $\text{np}(2\pi\sigma_p^2)$
do not depend on $\mathbf{w}$.

Therefore minimizing this negative log-likelihood
gives the same $\mathbf{w}$
as weighted least squares with
 $\beta_p = 1/\sigma_p^2$.

Three Equivalent Views

For Gaussian regression we can describe fitting three ways.

View	Goal
Error	minimize squared residuals
Geometry	project $\mathbf{y}$ onto the model space
Probability	maximize Gaussian likelihood

These are not three different algorithms.

They are three mathematical interpretations
of the same parameter estimate.

What did we learn?

- ◆ Weighted least squares gives different observations different influence through positive weights β_p .
- ◆ The weighted normal equations are $X^T W X \mathbf{w} = X^T W \mathbf{y}$.
- ◆ Weighted least squares is ordinary least squares after multiplying each row by the square root of its weight.
- ◆ Independent Gaussian noise with common variance leads to ordinary least squares by maximum likelihood.
- ◆ Unequal known Gaussian variances lead to weights $\beta_p = 1/\sigma_p^2$.
- ◆ Probabilistic regression describes both a predicted center and uncertainty around that prediction.



Next

- Regression predicts numerical responses.
- Probabilistic regression attaches a distribution to those numerical predictions.
- Next we change the kind of output: a finite set of categories.
- **1.14 - From Regression to Classification**