

Machine Learning

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

Previous Lecture



- ◆ Least squares minimizes the squared L^2 residual size; LAD minimizes the L^1 residual size.
- ◆ Squared loss gives increasingly large influence to large residuals;
absolute loss grows only linearly.
- ◆ For a constant model, least squares chooses the mean while LAD chooses a median.
- ◆ LAD is generally more robust to large vertical outliers than least squares.
- ◆ Absolute loss is nonsmooth at zero, so the normal-equation and ordinary
Newton machinery no longer applies directly.
- ◆ Choosing a loss function is part of choosing what “best fit” means.

§5.4 - Assessing a Regression Model

Last time:

different loss functions produced different notions of “best fit.”

Today:

once a model has been fit,
how do we decide whether it is any good?

One number is rarely enough.

Return to Our Regression Line

Study-hours data:

x	1	2	3	4	5
y	52	61	67	78	85

Least-squares model:

$$\hat{y} = 43.7 + 8.3x.$$

Predictions:

52.0, 60.3, 68.6, 76.9, 85.2.

Residuals:

0, -0.7, 1.6, -1.1, 0.2.

These residuals contain the information used by most regression metrics.

Sum of Squared Errors

Define

$$\text{SSE} = \sum_{p=1}^P (\hat{y}_p - y_p)^2.$$

For our model:

$$\begin{aligned}\text{SSE} &= 0^2 + (-0.7)^2 + (1.6)^2 + (-1.1)^2 + (0.2)^2 \\ &= 0 + 0.49 + 2.56 + 1.21 + 0.04 \\ &= 4.30.\end{aligned}$$

Smaller SSE:

predictions closer to the observed values.

But SSE grows automatically as the number of observations grows.

Mean Squared Error

Divide SSE by the number of observations:

$$\text{MSE} = \frac{1}{P} \sum_{p=1}^P (\hat{y}_p - y_p)^2.$$

For our data:

$$\begin{aligned}\text{MSE} &= \frac{4.30}{5} \\ &= 0.86.\end{aligned}$$

This is the same cost function we minimized.

Units:

if y is measured in score points,
MSE is measured in score-points squared.

Root Mean Squared Error

Take the square root:

$$\text{RMSE} = \sqrt{\text{MSE}}.$$

For our model:

$$\begin{aligned}\text{RMSE} &= \sqrt{0.86} \\ &\approx 0.927.\end{aligned}$$

Advantage:

RMSE has the same units as y .

Interpretation:

a typical squared-error scale is about
0.93 score points.

Mean Absolute Error

From Lesson 1.11:

$$\text{MAE} = \frac{1}{P} \sum_{p=1}^P |\hat{y}_p - y_p|.$$

For our residuals:

$$\text{MAE} = \frac{0+0.7+1.6+1.1+0.2}{5}$$

$$= \frac{3.6}{5}$$

$$= 0.72.$$

MAE also has the same units as y .

Compared with RMSE:
 MAE gives less extra weight to large residuals.

MSE, RMSE, and MAE

Metric	Formula	Units
SSE	$\sum r_p^2$	y^2
MSE	$\frac{1}{P} \sum r_p^2$	y^2
RMSE	$\sqrt{\frac{1}{P} \sum r_p^2}$	y
MAE	$\frac{1}{P} \sum r_p $	y

No single metric is universally “the” error.

The metric should reflect
 what kinds of prediction errors matter.

We Need a Baseline

Is
 MSE = 0.86
 good?

That depends on how much the observed y values vary.

A simple baseline model ignores x entirely.

Predict the same value for everyone:
 $\hat{y} = \bar{y}$.

For our data:
 $\bar{y} = \frac{343}{5}$
 = 68.6.

How much better is our regression line than this baseline?

Total Sum of Squares

Measure variation around the mean:
 $SST = \sum_{p=1}^P (y_p - \bar{y})^2.$

For our data:
 SST
 = $(52 - 68.6)^2 + (61 - 68.6)^2 + (67 - 68.6)^2$
 + $(78 - 68.6)^2 + (85 - 68.6)^2$

$$= 693.2.$$

SST is the squared error of the constant mean predictor.

The Coefficient of Determination

Compare model error with baseline error:

$$R^2 = 1 - \frac{SSE}{SST}.$$

For our fitted line:

$$R^2 = 1 - \frac{4.30}{693.2} \\ \approx 0.9938.$$

About 99.38% of the variation around the mean
is accounted for by the fitted values in this training sample.

Interpreting R^2

If

$R^2 = 1$:
perfect fit.

If

$R^2 = 0$:
same SSE as predicting the mean.

If

$R^2 < 0$:
worse than the mean baseline.

For an ordinary least-squares fit
with an intercept,
evaluated on its training data:
 $0 \leq R^2 \leq 1$.

On new data:
 R^2 can be negative.

R^2 Is Not an Error in the Units of the Problem

RMSE:
about 0.93 score points.

MAE:
about 0.72 score points.

R^2 :
dimensionless.

So they answer different questions.

RMSE / MAE:
how large are prediction errors?

R^2 :
how much better are we than the mean baseline?

Use them together.

A High R^2 Does Not Prove the Model Is Appropriate

A high R^2 does not imply:

- the relationship is causal;
- the relationship is truly linear;
- residuals have no pattern;
- predictions will remain accurate on new data.

A summary statistic can hide structure.

So:
look at the residuals.

Residual Plot

Plot residuals vertically against the input:
 x_p versus r_p .

For a useful linear model:
we hope to see residuals scattered around zero
without an obvious systematic pattern.

A residual plot can reveal problems
that a single MSE or R^2 value cannot.

Pattern 1: Curvature

Suppose residuals are

- negative for small x ,
- positive in the middle,
- negative again for large x .

A curved pattern remains.

Interpretation:
the linear model may be missing nonlinear structure.

The model systematically underpredicts in some regions
and overpredicts in others.

A low average error can still hide this pattern.

Pattern 2: Changing Spread

Suppose residuals are tightly clustered near zero for small x ,
but spread farther apart as x increases.

This is a **fan-shaped** residual plot.

Interpretation:
prediction variability changes with the input.

One global error number
does not describe every region equally well.

Pattern 3: An Unusual Observation

Suppose most residuals are near zero,
but one residual is very large.

Possible explanations:
measurement error;

unusual case;

missing feature;

model inadequacy.

Do not automatically delete the point.

First understand why it is unusual.

Same MSE, Different Residual Structure

Consider two residual vectors:

$$\mathbf{r}_A = (-2, -1, 0, 1, 2),$$

and

$$\mathbf{r}_B = (0, 0, \sqrt{10}, 0, 0).$$

For both:

$$\sum r_p^2 = 10,$$

so

$$\text{MSE} = 2.$$

But:
Model A spreads error across observations;

Model B has one large error.

Same MSE.

Very different diagnostic picture.

Training Error

So far we have calculated metrics on the same observations used to fit the model.

This is **training error**.

Training error answers:
how well does the fitted model match the data it saw?

Machine learning asks a harder question:
how well will the model predict observations it did **not** see?

That is a question about **generalization**.

Training Data and Test Data

A common strategy:
split the observations.

Training data:
used to choose the model parameters.

Test data:
held aside while fitting;

used later to assess prediction on unseen observations.

Then report, for example:
training RMSE,
test RMSE.

Large gap:
possible overfitting.

We will study this much more carefully later.

Why Training Error Usually Looks Better

The model parameters were chosen
specifically to reduce training loss.

Therefore the training observations had an advantage.

The test observations did not participate in fitting.

So:

- do not use low training error alone
- as evidence of strong predictive performance.

Fitting and assessment are different tasks.

Metrics in Python

Use the study-hours model:

```
x = np.array([1., 2., 3., 4., 5.])
y = np.array([52., 61., 67., 78., 85.])

w = np.array([43.7, 8.3])
y_pred = w[0] + w[1]*x

residuals = y_pred - y
```

Compute:

```
mse = np.mean(residuals**2)
rmse = np.sqrt(mse)
mae = np.mean(np.abs(residuals))

print(mse, rmse, mae)
```

Approximately:

```
0.86  0.92736  0.72
```

New Python idea:

np computes square roots.

Compute R^2 in Python

```
sse = np.sum((y_pred - y)**2)

y_bar = np.mean(y)
sst = np.sum((y - y_bar)**2)

r_squared = 1 - sse/sst
print(r_squared)
```

Output:

```
0.993796884...
```

The formula is exactly the one derived by hand.

Plot the Residuals in Python

```
plt.scatter(x, residuals)
plt.axhline(0)

plt.xlabel("Hours studied")
plt.ylabel("Residual")
plt.show()
```

New Python idea:
plt
draws the horizontal line
 $r = 0$.

Now we can inspect
where the model overpredicts and underpredicts.

A Practical Assessment Checklist

- After fitting a regression model:
- calculate an error metric in meaningful units;
 - compare with a simple baseline;
 - inspect residuals;
 - assess performance on data not used for fitting.

A model is not adequately assessed by saying only:
“the optimizer converged.”

Optimization finds parameters.

Assessment asks whether the resulting model is useful.

What did we learn?

- ◆ 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.

Next

- So far every observation has contributed equally to the regression loss.
- Sometimes observations have different reliability or importance.
- We can encode that information using weights.
- **1.13 - Weighted and Probabilistic Regression**