Review #1

SYS 6018 | Spring 2025

review-1.pdf

Contents

1 Supervised Learning

1.1 HW 1

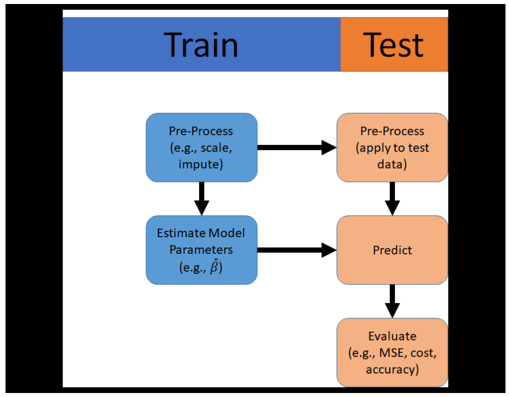
- The best predictive model is not always the true model.
  - Quadratic didn't always make best predictions. Why not?

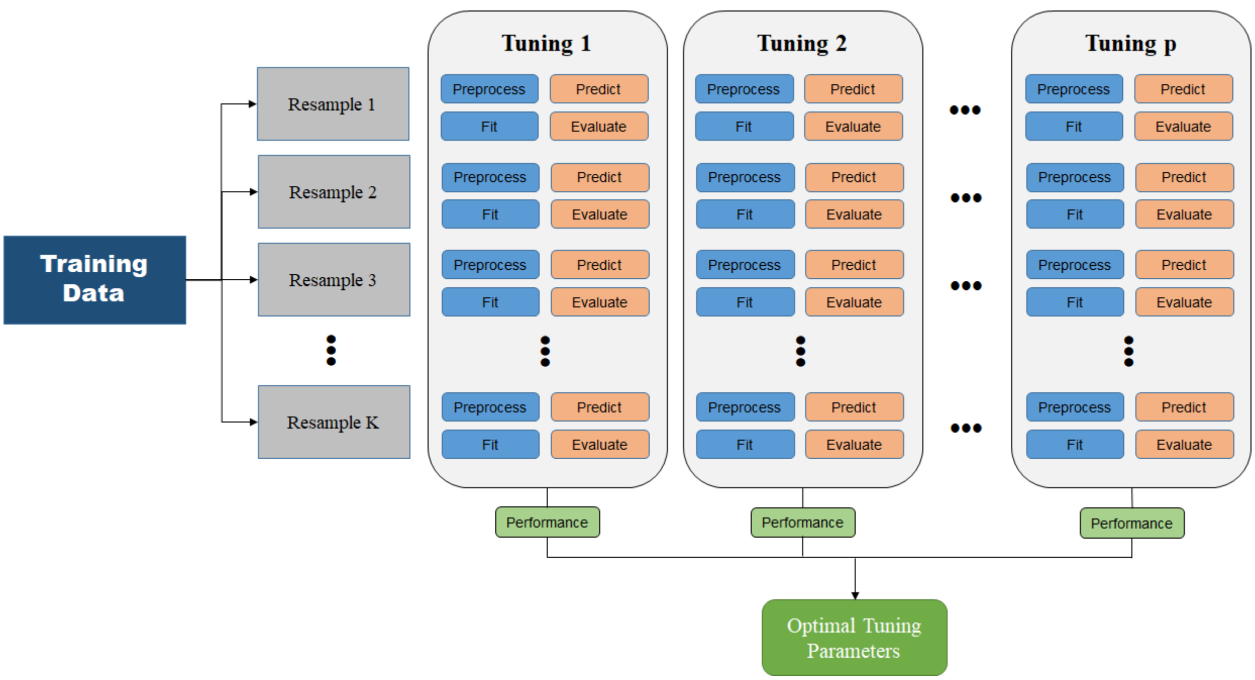
1.1.1 Questions

1. What is the Expected Prediction Error (EPE) (also known as Risk) and why do we care about it?
2. How is the EPE different from the training error (also known as empirical Risk)?
3. Under the squared error loss function, what is the optimal prediction? What about for the absolute error? Log loss?
4. How does model complexity/flexibility relate to the bias and variance of a predictive model?
5. What is overfitting? What is underfitting? How can they be prevented?
6. What are some ways to *increase* the complexity/flexibility of a predictive model? Ways to *decrease*?

2 Resampling

2.1 HW 2





2.2 Questions

1. What are some approaches to *estimate* the Expected Prediction Error (EPE)?
2. In a train/test split, what proportion of observations should go in test set? Why?
3. What is the primary purpose of the bootstrap method?
4. How much training data does K-fold CV use to estimate the model parameters?

5. How does the bootstrap method simulate new data?
6. What is the expected proportion of observations that will not appear in a bootstrap sample (out-of-bag)?
7. How can out-of-bag samples be used in model evaluation?
8. What are the advantages of using the bootstrap over traditional methods like deriving confidence intervals from normal distribution assumptions?
9. Explain the bias-variance tradeoff in the context of bootstrap aggregating (bagging)? How does the bootstrap help in reducing variance?
10. What is cross-validation? What is it used for?
11. What is difference between k-fold and monte-carlo cross-validation? What are the advantages of each?
12. What is difference between OOB and cross-validation?
13. What is stratified cross-validation and why is it useful?
14. What is nested cross-validation and how does it compare to train-validate-test splits?
15. What is the optimal K in K-fold cross-validation?
16. In comparing predictive models using cross-validation, is it OK if each model uses a different cross-validation folds?

3 Penalized Regression

3.1 HW 3

- Contest Results.
- How did reported performance match the actual performance on the test data?
- Why did the 1-SE approach not predict as well as $\lambda_{\min}$?

3.2 Questions

1. What is regularization (or penalization) in regression?
2. What is the bias-variance tradeoff using penalized estimation.
3. Compare the lasso, ridge, elastic net, and best subset penalties?
4. Compare the lambda min and one-standard error rule in penalized regression.
5. What is one way to compare predictions of two models on a test set?

4 Tree-based methods

4.1 HW 4

- Random Forest Tuning

4.2 Questions

1. Explain how CART (classification and regression trees) work?
2. In a classification/probability tree, how are splits made?
3. In a regression tree, how are splits made?
4. In a classification/probability tree, what are the predictions made in the leaf nodes?
5. How are trees similar to nearest neighbor models?
6. What are the tuning parameters in CART? How do they impact bias and variance?
7. How does the OOB error work in Random Forest? How does the number of trees impact the uncertainty in this estimate? What is an advantage of OOB over cross-validation in RF?
8. Why do I not suggest tuning the number of trees in Random Forest?
9. Why are Random Forests called an *ensemble model*?
10. Why is combining predictions from multiple trees expected to improve performance?
11. What are the main tuning parameters in Random Forest. Do they impact bias, variance, or something else?

5 SVM

1. How are SVMs similar to Logistic Regression?
2. What are the “kernels” in SVM?
3. What are “support vectors” in SVM?
4. What is the loss function used by SVMs? What is the penalty?
5. What is one way to convert the output from SVM into a probability?
6. Wny does probability calibration for SVM not expected to work well?
7. How does the Radial Basis Function (RBF) kernel work in SVM?
8. Suppose you have a large dataset with millions of features. How would you optimize SVM to handle this efficiently?
9. How would you choose the best kernel for your SVM model?
10. What are some advantages and disadvantages of using SVM compared to other classifiers like logistic regression or random forests?

6 Classification

6.1 HW 5

6.1.1 Contest Part 1 Results

6.1.2 Contest Part 2 Results

6.2 Questions

1. What is the logit function?
2. What is the standard loss function used in logistic regression?
3. Explain how logistic regression make probability outputs?
4. How can hard classifications be made in logistic regression?
5. How can you assess the performance of a logistic regression model?
6. What are some methods to handle class imbalance in logistic regression?
7. What is the maximum likelihood estimation, and how is it used in logistic regression?
8. What is the difference between accuracy, precision, recall, and F1 score?
9. What is a confusion matrix, and how is it used in the evaluation of classification models?
10. Suppose your logistic regression model has high accuracy but poor recall. How would you improve it?
11. How would undersampling influence the predictive performance of a classification model?
12. Can ROC curves and AUC tell you which observations are predicted poorly?
13. How can you tell which types of observations are predicted poorly?
14. How should you choose the classification threshold if you have to make a hard decision?
15. Why do I say it may be unethical for a predictive model to make a hard classification?
16. How does class unbalance influence the quality of a predictive model? Which types of models are most impacted by class unbalance?
17. Should anything be done if there is class unbalance?

7 Calibration Curves

The textbook *An Introduction to Statistical Learning (ISL)* has a description of a simulated credit card default dataset. The interest is on predicting whether an individual will default on their credit card payment.

```
data(Default, package="ISLR")

#: Create binary column (y)
Default = Default %>% mutate(y = if_else(default == "Yes", 1L, 0L))
```

The variables are:

- *outcome variable* is categorical (factor) `Yes` and `No`, (`default`)
- the categorical (factor) variable (`student`) is either `Yes` or `No`
- the average balance a customer has after making their monthly payment (`balance`)
- the customer's income (`income`)

```
set.seed(11)
Default %>% slice_sample(n=6)
```

default	student	balance	income	y
No	No	396.5	41970	0
No	No	913.6	46907	0
No	Yes	561.4	21747	0
Yes	Yes	1889.3	22652	1
No	No	491.0	37836	0
No	Yes	282.2	19809	0

A risk model is said to be *calibrated* if the predicted probabilities are equal to the true risk (probabilities).

$$\Pr(Y = 1 \mid \hat{p} = p) = p \qquad \text{for all } p$$

Create train/test split

```
#: train/test split
set.seed(2019)
test = sample(nrow(Default), size=2000)
train = -test
```

Fit logistic regression model to training data

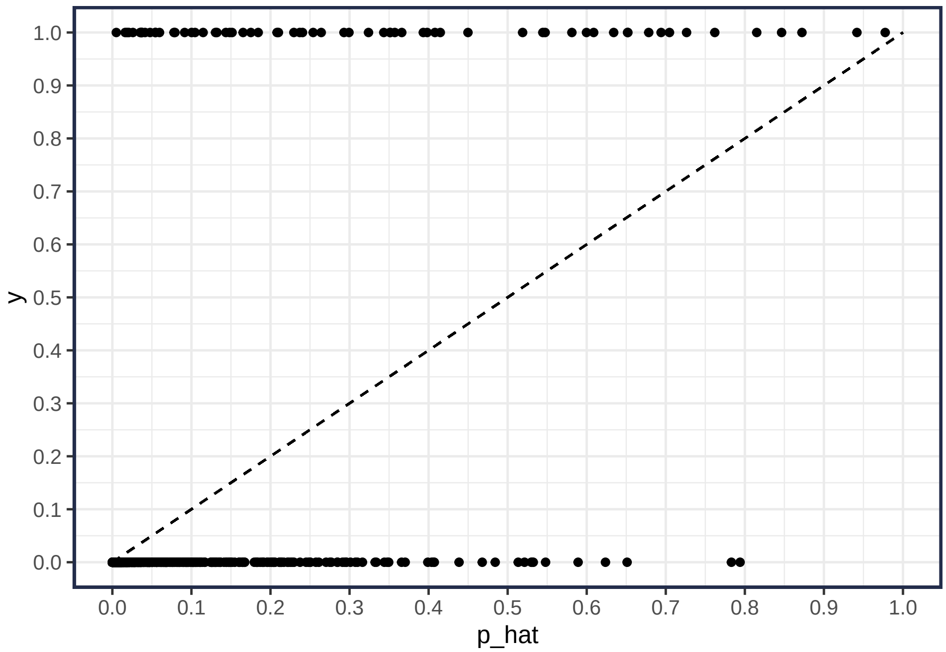
```
#: fit logistic regression on training data
fit.lm = glm(y~student + balance + income, family='binomial',
             data=Default[train, ])
```

Make predictions on test data

```
p_hat = predict(fit.lm, Default[test,], type="response")
preds_test = tibble(
  y = Default$y[test],
  student = Default$student[test],
  p_hat = p_hat
)
```

```
plt = preds_test %>%
  ggplot(aes(p_hat, y)) + geom_point() +
  scale_x_continuous(breaks = seq(0, 1, by=.1)) +
```

```r
  scale_y_continuous(breaks = seq(0, 1, by=.1)) +
  coord_cartesian(xlim = c(0,1), ylim=c(0,1)) +
  geom_segment(x = 0, xend = 1, y = 0, yend = 1, lty=2) # perfect calibration
plt
```

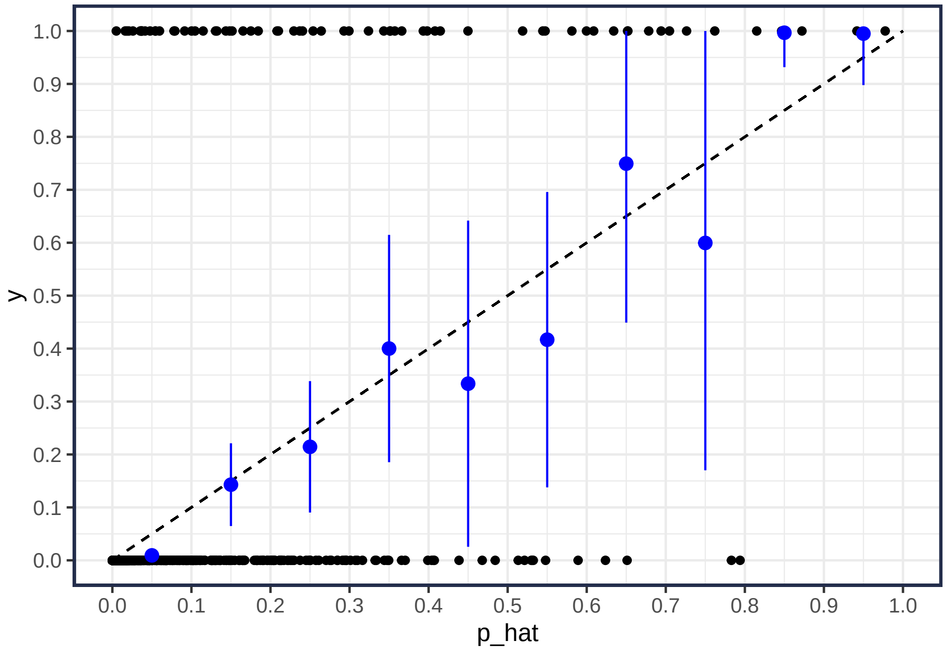


Create bins along the x-axis ($\hat{p}$) and calculate the mean response in each bin. Using Laplace smoothing to avoid extreme $\{0, 1\}$ estimates.

```r
bks = seq(0, 1, by = .10)
mids = bks[-1] - diff(bks)/2
binned_data = preds_test %>%
  mutate(
    p_hat_bin = cut(p_hat, breaks = bks, include.lowest = TRUE),
    midpoint = mids[as.integer(p_hat_bin)]
  ) %>%
  group_by(midpoint) %>%
  summarize(
    n = n(),
    n1 = sum(y == 1) + .01, # add .01 defaults to each bin
    n0 = sum(y == 0) + .01, # add .01 non-defaults to each bin
    p = n1 / (n0 + n1),
    se = sqrt(p*(1-p)/n),
    upper = pmin(p + 1.96*se, 1),
    lower = pmax(p - 1.96*se, 0)
  )
binned_data
#> # A tibble: 10 x 8
#>   midpoint     n    n1      n0       p      se  upper   lower
#>      <dbl> <int> <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>
#> 1     0.05  1822 17.0  1805.   0.00934 0.00225 0.0138 0.00492
#> 2     0.15    77 11.0    66.0  0.143   0.0399  0.221  0.0648
#> 3     0.25    42  9.01   33.0  0.214   0.0633  0.339  0.0903
#> 4     0.35    20  8.01   12.0  0.400   0.110   0.615  0.185
#> 5     0.45     9  3.01    6.01 0.334   0.157   0.642  0.0256
#> 6     0.55    12  5.01    7.01 0.417   0.142   0.696  0.138
#> # i 4 more rows
```

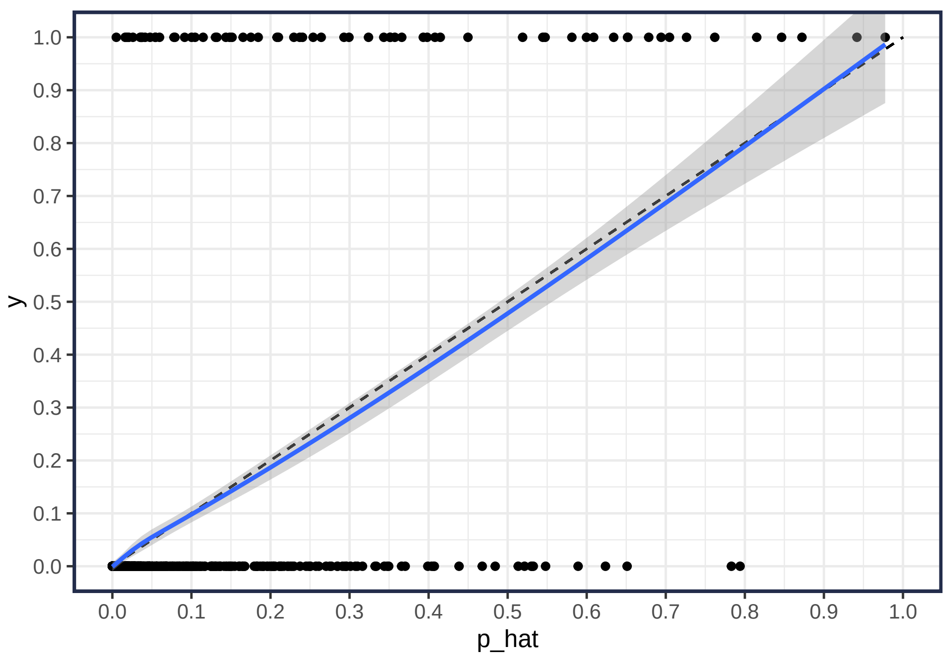
Plot binned estimates.

```
plt +
  geom_pointrange(data = binned_data,
                  aes(midpoint, y=p, ymin=lower, ymax=upper),
                  color = "blue")
```



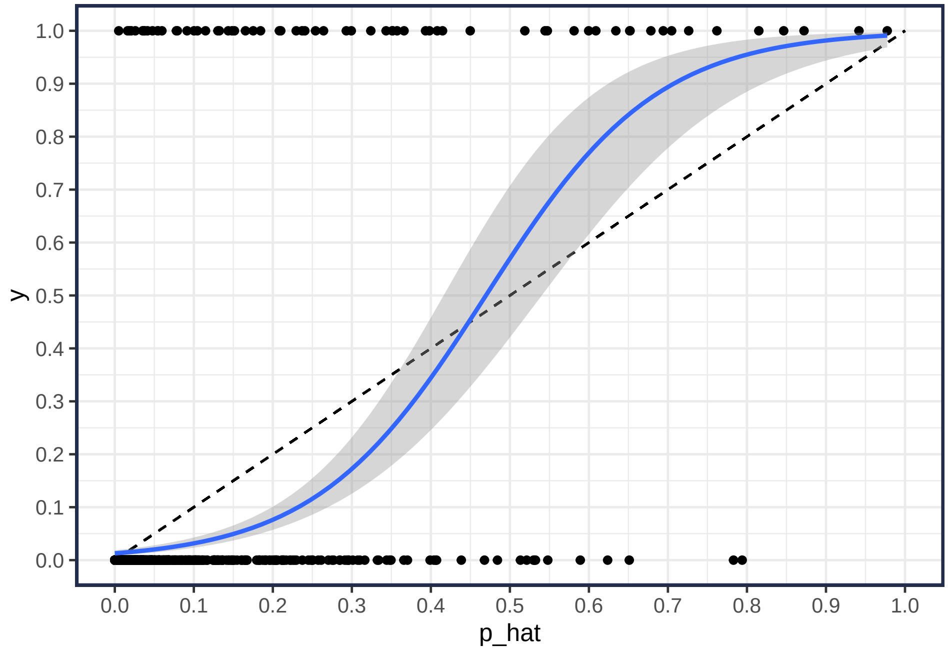
We could have instead added smooth line fit (predictor variable is $\hat{p}$, outcome variable is y). Note that this implements linear regression (squared error loss).

```
plt + geom_smooth()
```



A better way that incorporates the uncertainty that varies with $\hat{p}$ is to use logistic regression. If we try the add directly into `geom_smooth()` it doesn't looks quite right, why?

```
plt + geom_smooth(method = "glm", method.args = list(family = "binomial"))
```



Think of the structure of logistic regression - the linear component captures the *logit* of p (what we referred to as γ in a previous class). I.e.,

$$\text{logit } p(x) = \beta_0 + \beta_1 \hat{p}(x)$$

but we don't want this!

Rather, something like this is what we want

$$\text{logit } p(x) = \beta_0 + \beta_1 \hat{p}(x) + \text{logit } \hat{p}(x)$$

fit on a hold-out set, and check how far β_0 and β_1 are from 0.

```
preds_test %>%
  mutate(gamma = log(p_hat) - log(1-p_hat)) %>%
  glm(y ~ p_hat + offset(gamma), family = "binomial", data = .) %>%
  broom::tidy()
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic p.value
#>   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
#> 1 (Intercept)  0.00981     0.221    0.0444   0.965
#> 2 p_hat       -0.187       0.720   -0.259    0.795
```

Or examine non-linear deviations with B-splines:

```
library(splines)
smooth_fit = preds_test %>%
  mutate(gamma = log(p_hat) - log(1-p_hat)) %>%
  glm(y ~ splines::bs(p_hat) + offset(gamma),
      family = "binomial", data = .)
```

```
smooth_fit %>% broom::tidy()
#> # A tibble: 4 x 5
#>   term                    estimate std.error statistic p.value
#>   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>
#> 1 (Intercept)               0.0224     0.348    0.0645   0.949
```

```
#> 2 splines::bs(p_hat)1    0.0241     1.53     0.0158    0.987
#> 3 splines::bs(p_hat)2   -0.677      2.20    -0.308     0.758
#> 4 splines::bs(p_hat)3    0.558      2.58     0.216     0.829
```

7.1 Questions

1. What does it mean for a binary classification model to be calibrated? How does this differ from standard measures of predictive performance?
2. Describe a calibration plot (reliability diagram). Explain how it is constructed.
3. Consider two models, A and B, with identical AUC-ROC scores of 0.85. Model A has a Brier score of 0.10, while Model B has a Brier score of 0.18. What does this suggest about the calibration of each model? Why is it possible for two models with the same AUC to have different calibration properties?
4. You generate a calibration plot (reliability diagram) from predictive probabilities. The curve consistently lies above the diagonal line. What does this indicate about the model's probability predictions? How would this affect decision-making based on the model's outputs?
5. Compare and contrast Platt scaling and isotonic regression for probability calibration in SVM models. In what scenarios is one method preferable over the other? How do these methods handle non-monotonic calibration curves?
6. What are some methods to assess the calibration of a binary prediction model?
7. Describe a model-based statistical test to assess if predictions are calibrated.
8. Suppose you train a logistic regression model with L2 regularization (ridge regression). How does increasing the regularization strength affect model calibration? Would you expect over-regularization to lead to underconfident or overconfident predictions?
9. Is a validation/test set comprising only 10% of the total data (e.g., a 90-10 split) sufficient to reliably assess calibration in a binary classification model? Why or why not? What factors influence whether this proportion is adequate?