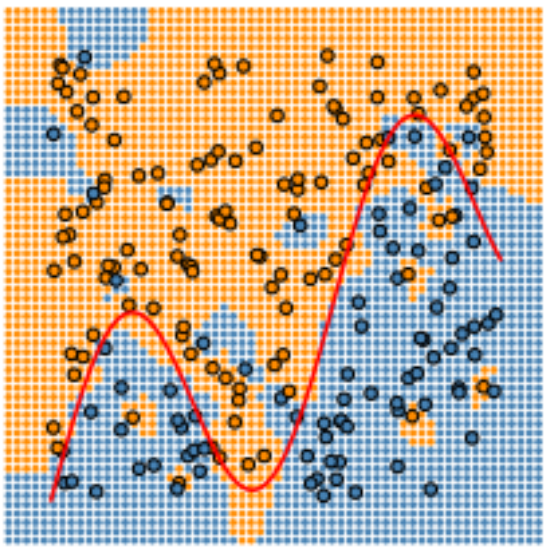
Model Selection

Nan Ye

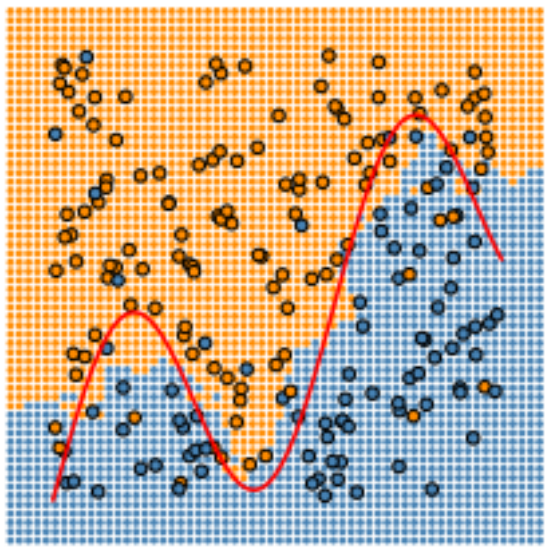
School of Mathematics and Physics
The University of Queensland

Recall

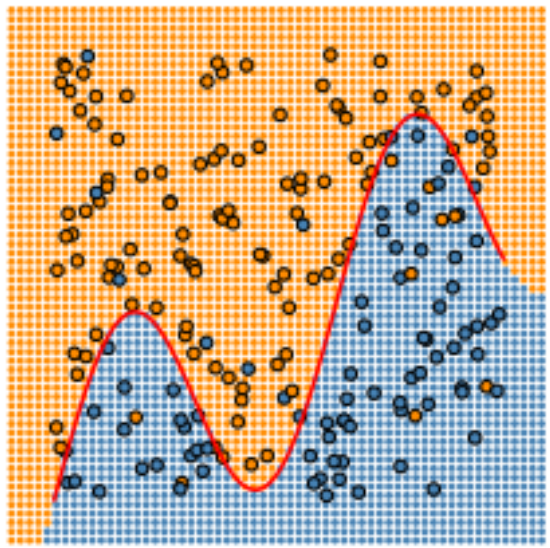
1-NN classifier



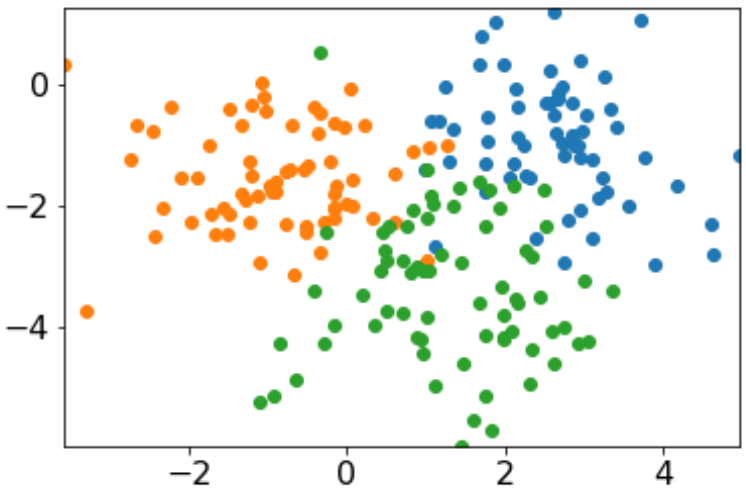
10-NN classifier



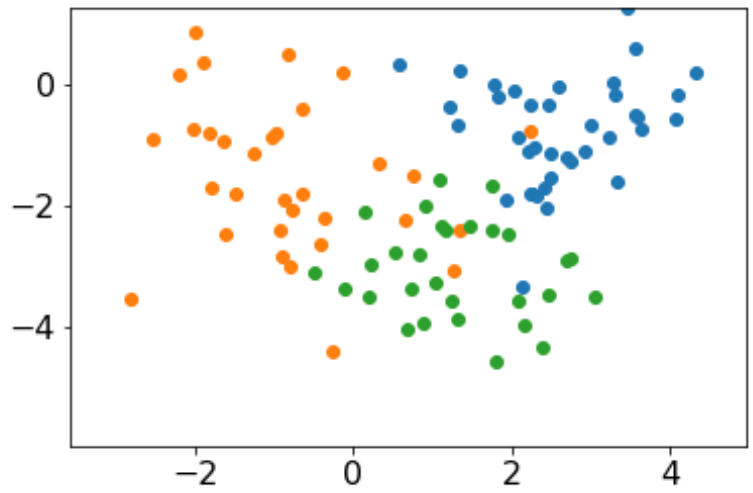
Bayes optimal classifier



For kNN, we need to choose the right k so that it works well.

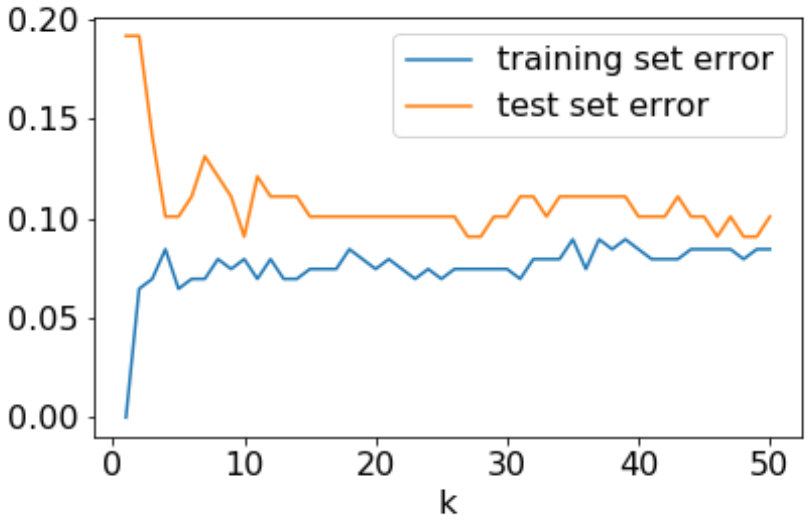


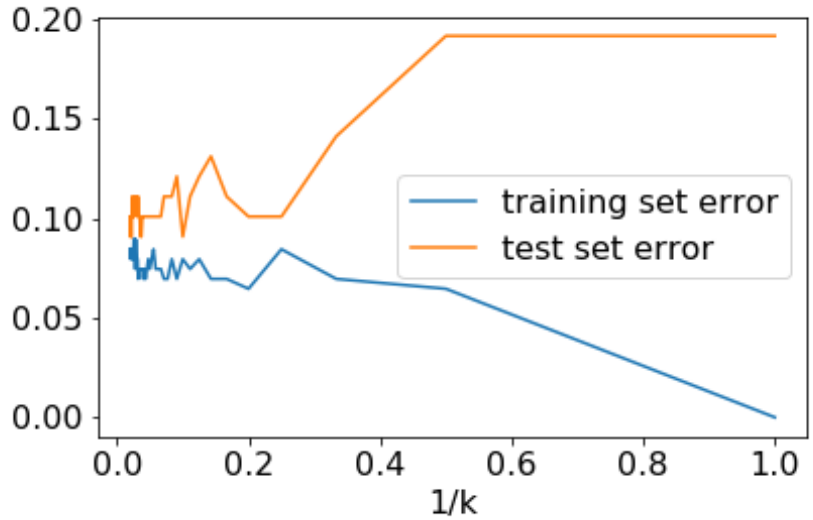
(a) Training data



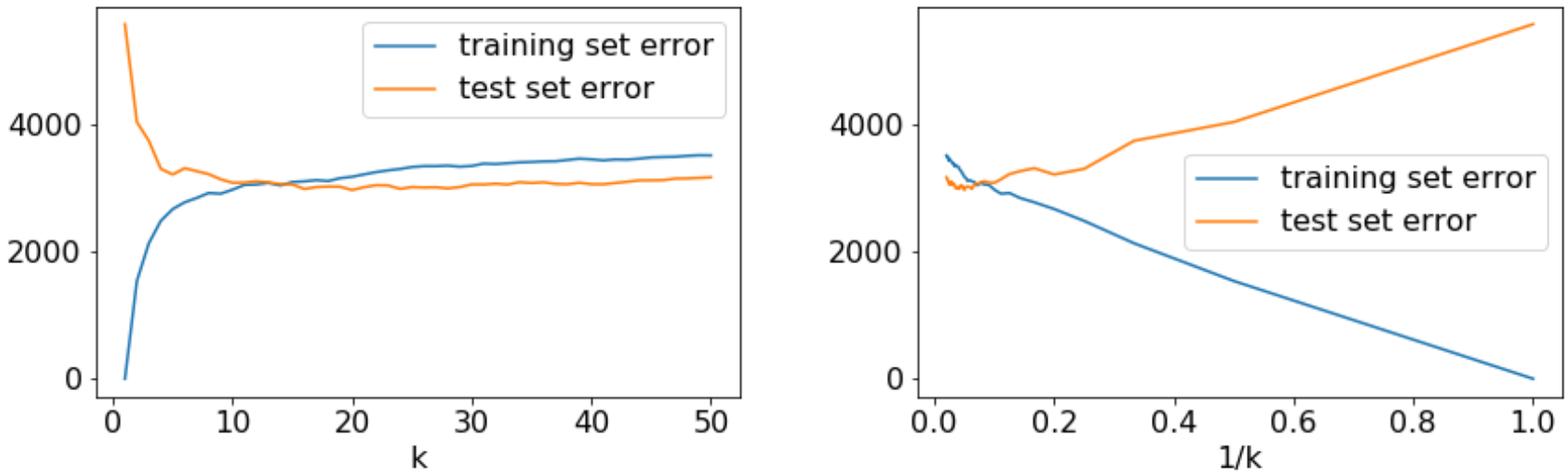
(b) Test data

Training and test set error of kNN classifier





kNN regression



Mean squared error for kNN regression on the diabetes dataset

Machine Learning A Cappella - Overfitting Thriller!

Udacity

Subscribe 45,724

31,516

Model Selection

- Model selection is concerned with estimating the performance of different models in order to choose the best one.
- We would hope that we can pick a model that can help us to drive the predictive error to zero.
- This is generally not possible.
  - A simple model will not pick up the regularities in data.
  - A complex model may pick up too much irregularities in data.
- This can be described more precisely by the bias-variance tradeoff.

Bias-Variance Tradeoff

- The bias and variance are two components of the predictive error.
- In the case of regression, consider how a learning algorithm performs on an input $\mathbf{x}$.
  - The output Y follows the conditional distribution $P(Y \mid \mathbf{x})$.
  - The predicted value Y' can be considered a random function of the training set.
  - We are interested in the expected prediction error $\mathbb{E}((Y' - Y)^2)$, where expectation is taken wrt to both Y and the random training set (which Y' depends on).
  - The expected prediction error is a property of the model class.

Bias-variance decomposition

$$\overbrace{\mathbb{E}\big((Y' - Y)^2\big)}^{\text{expected prediction error}} = \overbrace{\mathbb{E}\big((Y' - \mathbb{E}(Y'))^2\big)}^{\text{variance}} + \overbrace{\big(\mathbb{E}(Y') - \mathbb{E}(Y)\big)^2}^{\text{bias (squared)}} + \overbrace{\mathbb{E}\big((Y - \mathbb{E}(Y))^2\big)}^{\text{irreducible noise}},$$

Proof. Expand the RHS and simplify.

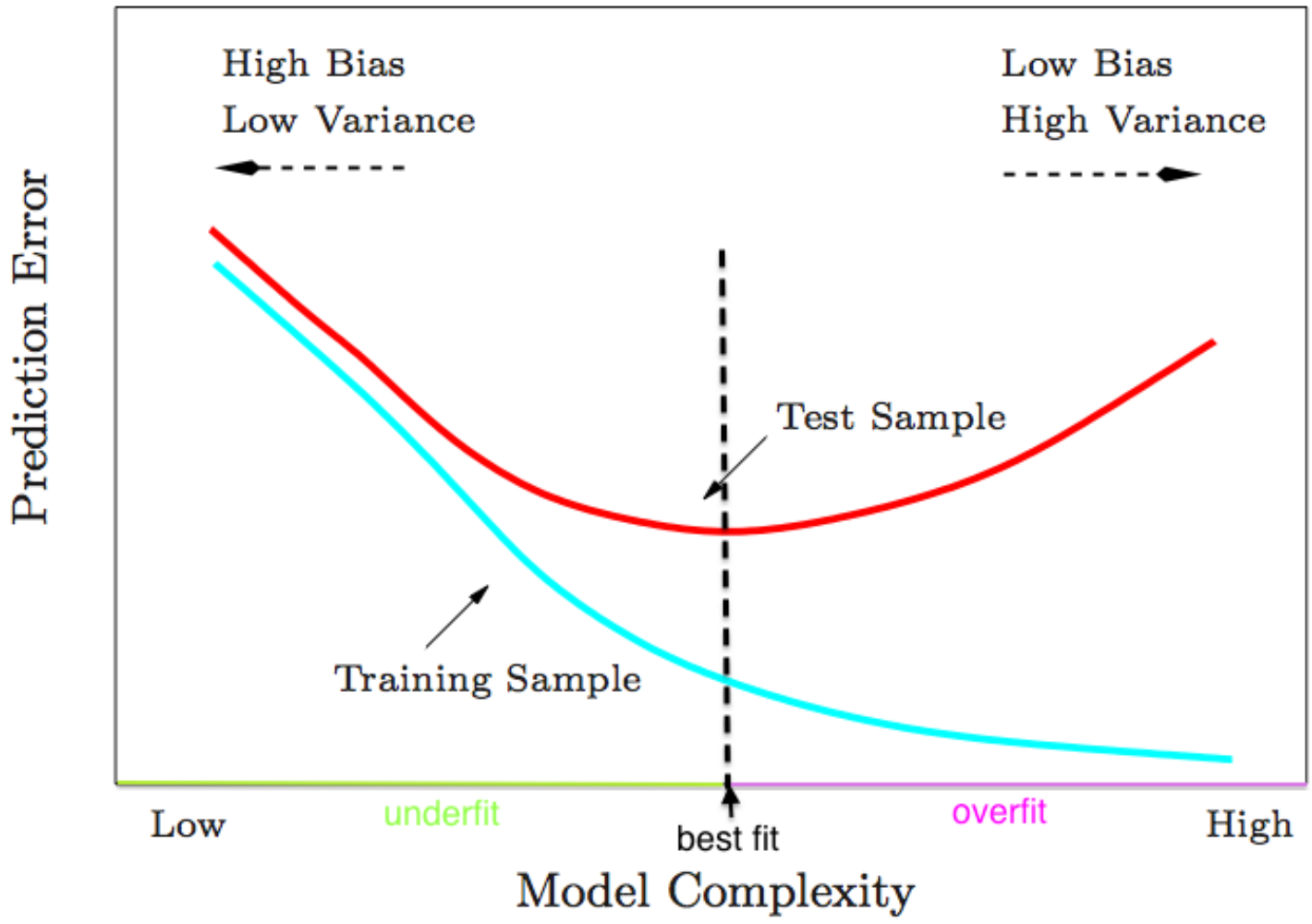
Bias-variance decomposition

$$\overbrace{\mathbb{E}\big((Y' - Y)^2\big)}^{\text{expected prediction error}} = \overbrace{\mathbb{E}\big((Y' - \mathbb{E}(Y'))^2\big)}^{\text{variance}} + \overbrace{\big(\mathbb{E}(Y') - \mathbb{E}(Y)\big)^2}^{\text{bias (squared)}} + \overbrace{\mathbb{E}\big((Y - \mathbb{E}(Y))^2\big)}^{\text{irreducible noise}},$$

Proof. Expand the RHS and simplify.

Bias-variance tradeoff

In general, as model complexity increases (i.e., the hypothesis becomes more complex), variance tends to increase, and bias tends to decrease.


High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Training Sample
Low
underfit
best fit
overfit
High
Model Complexity

Bias-variance Tradeoff in kNN (optional)

Assumption

- Suppose $Y \mid X \sim N(f(X), \sigma^2)$ for some function f and some fixed σ. Put it in another way,

$$Y = f(X) + \epsilon,$$

  where $\epsilon \sim N(0, \sigma^2)$.
- In addition, we consider the simpler case where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are fixed.

Complexity measure for kNN

- We can consider $\frac{1}{k}$ as a complexity measure for the kNN model.
- When $\frac{1}{k}$ is small, a kNN model is closer to a constant (the average of all outputs), and thus the model is simpler.
- On the other hand, when $\frac{1}{k}$ is large, a kNN model is likely to be very complex function, and thus the model is more complex.

Bias and variance

- Let Y be the true value at $\mathbf{x}$, then $\mathbb{E}\, Y = f(\mathbf{x})$.
- kNN predicts the value $Y' = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i$ for $\mathbf{x}$.
- With some calculation, we have

$$\textit{bias} = \mathbb{E}(Y') - \mathbb{E}(Y) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} f(\mathbf{x}_i) - f(\mathbf{x}),$$

$$\textit{variance} = \mathbb{E}((Y' - \mathbb{E}(Y'))^2) = \sigma^2 / k.$$

Bias-variance trade-off

As $\frac{1}{k}$ increases (or as model complexity increases), bias is likely to decrease, and variance increases.

Model Selection Techniques

- We often need to consider the following problems
  - **(a)** Given several different classes of models (e.g. logistic regression, SVMs), how do we choose the best class?
  - **(b)** Given a model with some tunable hyperparameters (e.g. the regularization constant in ridge regression), how do we choose the best hyperparameters?

- In the next few slides, we will think of (b) as a special case of (a) by taking each hyperparameter configuration as defining a class of models.
- We can phrase our problem as follows: given m classes of models $\mathcal{M}_1, \ldots, \mathcal{M}_m$, how do we choose the best model?

  here each $\mathcal{M}_i$ is a set of *concrete* models, such as functions, or probability distributions

Using a validation set

- Split the data into a training set $\mathcal{T}$, a validation set $\mathcal{D}$.
- For each $\mathcal{M}_i$, train a model on $\mathcal{T}$, and test it on $\mathcal{D}$.
- Choose the model with best validation set performance.

Remarks

- A lot of data is needed, while the amount may be limited.
- The validation set performance is not a good indicator of a model's ability to generalise (make good predictions on new inputs).
- The generalization performance of a model need to be assessed on a *separate* test set.

K-fold cross validation

- Split the training data into K folds.
- For each $\mathcal{M}_i$, train K models, with each trained on $K-1$ folds and tested on the remaining fold.
- Choose the parameter with best average performance.

Computationally more expensive than using a development set.

Variants of cross validation

- Leave-one-out cross validation is n-fold cross validation (that is, each fold has exactly one example).
- For data that can be divided into several groups, stratified cross validation aims to ensure the folds have similar group proportions.
- In practice, 10-fold stratefied cross validation is recommended, provided that computation time is not an issue, and the dataset is reasonably large.

More on Model Selection

- Two common approaches
  - Analytically approximate the validation step
    - AIC, BIC, MDL
  - Efficient sample re-use
    - Cross validation, bootstrap

What You Need to Know...

- Goodness of fit $\neq$ predictive performance, overfitting
- Bias-variance tradeoff
- Two methods: development set and cross-validation.