Class 1: Course Overview & Introduction to Statistical Learning

MSA 8150: Machine Learning for Analytics

Alireza Aghasi

Institute for Insight, Georgia State University

About the Course

Instructor

- **Instructor:** Alireza Aghasi
- **Office:** Room 542, Buckhead Center
  Room 405B, Downtown Campus
- **Office Hours:** Wednesdays 6:00-7:15p (@ Buckhead Center)
  (starting second week)
- **Email:** aaghasi@gsu.edu

– What if you cannot make it to the office hours during the times above?

Will discuss this in a few minutes!

Course Related

- **Course Material Available at:** iCollege http://icollege.gsu.edu
- **Teaching Assistant:** TBA (@student.gsu.edu)
    - Issues related to the homework grades should be discussed with the TA
    - If an issue was not resolved, then it could be discussed with the instructor during the office hours
- **Prerequisite:** Some level of exposure to statistics, basic calculus and linear algebra
- **Programming:** R (Mainly), Python, MATLAB
- Please note that this is not a programming course and students are required to develop the required programming skills on their own

Text Book

- James, G., Witten, D., Hastie, T., Tibshirani, R., "An Introduction to Statistical Learning: with Applications in R", Springer, 2013 **(main text)**. URL (E-book): https://goo.gl/8NYEo4
- Hastie, T., Tibshirani, R., Friedman, J., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Second Edition, Springer, 2009 **(supplemental text)** URL (E-book): https://goo.gl/xgr63x

  : Let's be green and help the environment!

Additional Reading:

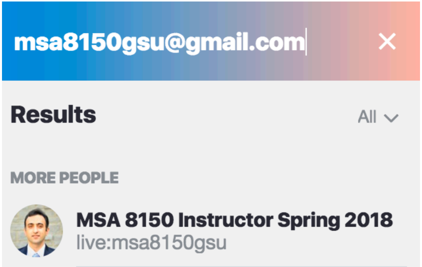
- Bishop, C., "Pattern Recognition and Machine Learning", Springer, 2006. URL: https://goo.gl/56GFVv
- Mitchell, T.M., "Machine Learning", McGraw-Hill, 1997. URL: https://goo.gl/HrBDtK

Communications with the Instructor

- The majority of questions and concerns can be addressed during the office hours

What if someone cannot make it?

- You may remotely communicate with the instructor during the office hours (only) via **Skype**:



- Note: you may need to wait if the instructor is already assisting others

Graduate Course: Yet a New Collaborative Experience!

- The primary goal of the course is a **better learning experience** in a **collaborative environment**
- This is a large class, please avoid unnecessary communications/ emails and be more research focused!
  - `‘‘Hi Prof ..., what size notebook do I need for the class?’’`
  - `‘‘Do you recommend taking notes on a laptop or a writing pad?’’`
  - `‘‘How hard is programming?  What language should I pick?’’`
  - `‘‘I have written a piece of code and it keeps throwing errors, can you read the code and find out why?’’`
- No emails about a problem that can be resolved during the office hour
- Students are encouraged to use **PIAZZA** for a more collaborative environment

PIAZZA

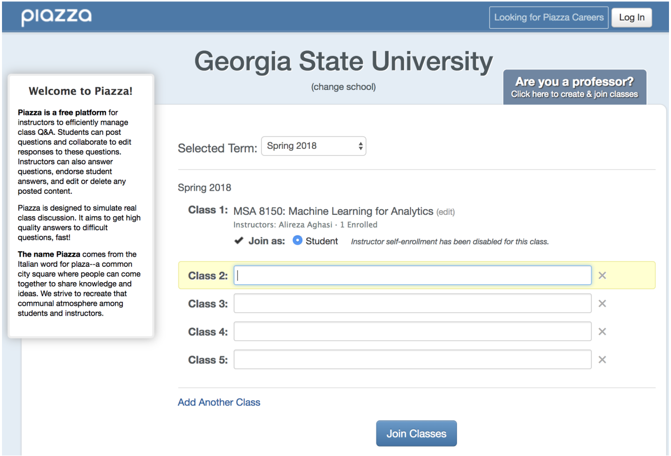
- We have set up a Piazza account for the course

What is Piazza?

- Piazza is a Q&A platform designed to get you answers from classmates and instructors fast.
- You would need to sign up with your GSU account

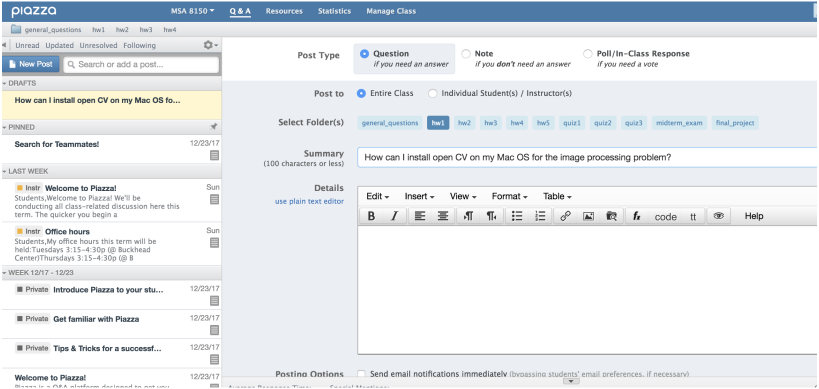
PIAZZA

Sign up page: piazza.com/gsu/spring2019/msa8150



PIAZZA

You can select a **topic** such as `general question`, `hw1`, `exam`, `...` and post a question



PIAZZA

- The majority of questions you might have about the course, homework, resources can be addressed by your colleagues, so try Piazza first!
- A wide range of questions can be asked on Piazza (of course as long as it is **course related** and **asked professionally**)
- Top **10%** of the PIAZZA active users who are professionally the most helpful to their colleagues **will be awarded extra credit at the end of the semester**

PIAZZA Rules & Guidelines (Important)

While the use of Piazza is included in the course as a tool to construct a collaborative environment, there are some strict rules that need to be followed:

- Always be professional in asking/answering questions. Penalties apply to the violators.
- You are **not allowed** to explicitly **ask a homework question before 8:00 AM of the day after the due date (i.e., 8 AM Thursday)**.
- Only general questions related to the homework or problems of similar type can be posted. Violators will receive zero credit for the corresponding homework.
- You are **not allowed** to post the solution to any of the homework questions **before 8:00 AM of the day after the due date (i.e., 8 AM Thursday)**. Violators will receive zero credit for the corresponding homework.

PIAZZA Rules & Guidelines (Important)

- Students are free to ask questions or discuss problems prior to a quiz or the midterm exam
- Students are **not allowed** to post **anything related to the material presented in the quiz (or the midterm exam) before 8:00 AM of the Thursday following the quiz (or the midterm exam)**. Serious penalties apply to the violators.
- Details on the homework posting and due dates will be presented later in the slides

Grading

There are multiple items that contribute to the final course grade as listed below. There would be **no final exam** and instead a final project is evaluated.

Item	**Percentages**
Homework (4-5 Assignments)	20%
In-Class Quizzes (3 quizzes, top 2 counted, each 5%)	10%
Midterm Exam	25%
Knowledge Assessment Test	5%
Final Project	35%
Attendance (see below)	5%

The final grade conversion is based on the following table:

A+	A	A-	B+	B	B-	C+	C	C-	D	F
≥ 97	≥ 90	≥ 87	≥ 83	≥ 80	≥ 77	≥ 73	≥ 70	≥ 67	≥ 60	<60

Teaching Strategy

- The class will be taught in an interactive way
  - Students are encouraged to take part in the class discussions
  - Students are invited to present their ideas in solving problems when presenting examples
- Each session the course material will be taught from the slides
- Students are strongly encouraged to take notes in the class, since many examples and topics will be interactively discussed in the class
- In taking notes, you do not need to rewrite what is already in the slides, **just write down the slide number and add the related notes next to it**

Homework

- Homework will be assigned on a biweekly basis (approximately)
  - Homework will be turned in at the beginning of the lecture
  - You will have approximately 10 days to work on each homework
  - Homework will be posted around a Sunday (or a Monday) and will be collected on the Wednesday of the week after
- Start working on the homework early
  - While you will have the office hour option during this time, the last ones are right before the due date which may not give you enough time to make changes
- Late homework **is not accepted** and will receive zero credit. If you cannot make it to the class for any reason:
  - either submit it electronically before the beginning of the lecture
  - ask a classmate to hand it in on your behalf

Homework

- Each student **must write up and turn in their own solutions**.
- Students copying from their classmates or from any other resources will receive a zero score.
- **IMPORTANT** If you solve a question together with another colleague, each need to write up your own solution and **need to list the name of people who you discussed the problem with on the first page of your turned in material**
- Effectively, homework is worth much more than 20% of your grade.
- The in-class quizzes are often taken immediately after turning in the homework and heavily correlated with the material turned in.
- It is extremely unlikely that ones does well on the exam/quizzes without putting enough effort into the homework

Attendance Policy

- **Lecture attendance is mandatory and will count towards your grade**
- sign in sheet (starting in week 2) will be passed around at every lecture
- sign next to your name (only)

Quizzes

- There will be 3 in-class quizzes
- The date of the quiz will be announced ahead of the time
- Most likely, they will be immediately after turning in a homework assignment and heavily correlated with the content of the homework
- Among the 3 quizzes the top 2 will counted towards your grade
- **There will be no make-up for the quizzes**

Midterm Exam

- The exam and all quizzes are closed-book and closed-notes
- Unless you are explicitly told that a calculator is allowed on a quiz or exam, there should not be any calculator within your reach during a quiz or an exam
- Make-up exams will be scheduled only in case of unavoidable emergencies and after the department's approval
- Personal business, such as interviews and travel arrangements do not warrant a make-up exam or an incomplete grade
- In case of an unavoidable emergency affecting your exam, you should contact the instructor, immediately

Midterm Exam

- All course participants are expected and required to abide by the Georgia State University Honor Code
- For details, see the University policy on academic honesty (section 409):
  http://www2.gsu.edu/ wwwfhb/sec409.html
- Please familiarize yourself with the code, and use it to guide your conduct
- You must do your own work in all the homework assignments, quizzes and exams
- Any form of academic dishonesty, such as plagiarism, can result in serious consequences and will be reported to the department administration

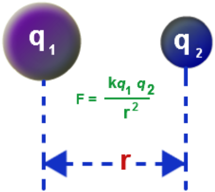
Knowledge Assessment Test & Final Project

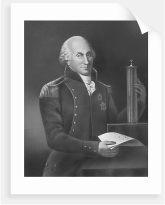
- Knowledge Assessment Test (KAT) is a comprehensive quiz (mainly conceptual questions) taken towards the end of the semester. It counts towards 5% of your total score
- Instructions and guidelines about the final project will be given around the middle of the semester, once students are familiar with the most fundamental tools in ML.

Introduction to Machine Learning

Yet Looking for Tools to Model

- We have all seen simple physics equations, discovered hundreds of years ago
- Since then we have been curious about how systems work and to predict their behavior before constructing them





- Back then, many derivations were based on experiments
- For instance deriving an equation like $F = 8.99 \times 10^9 \frac{q_1 q_2}{r^2}$ could involve fixing all the parameters except one and evaluating the response behavior
- In this simple example q_1, q_2, r are variables and F is a response

Yet Looking for Tools to Model

- What has changed?
  - Back then, we did not have accurate measurement systems but now we have very accurate measurement sensors for pretty much all physical phenomena
  - Back then, we were only able to collect limited data, now we are able to collect huge data
  - Back then, we did not have computational resources, now we have supercomputers
  - Back then, we alway liked more compact models for an easier application, now we can have super-complex models and ask the computers to apply them
- **Then why not model large and complex systems??!!**

Yet Looking for Tools to Model

– Models that can predict the weather, the stock market value, political interactions, ...

– In our century, having more data means more power to predict!

More Examples: Spam Detection

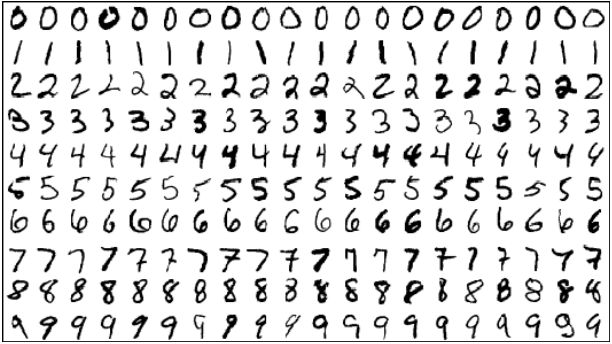
- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as spam or email
- goal: build a customized spam filter
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between **spam** and **email**

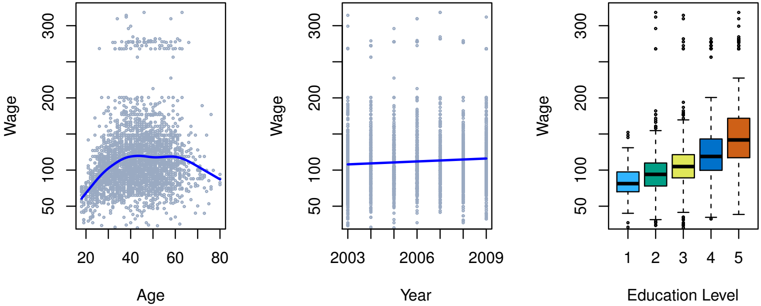
More Examples: Optical Character Recognition

identify hand-written characters based on the combination of all pixel values




More Examples: Salary Prediction

– Establish the relationship between salary and demographic variables in population survey data

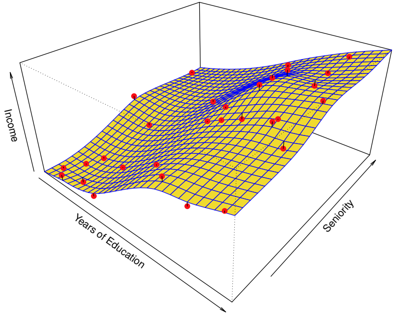


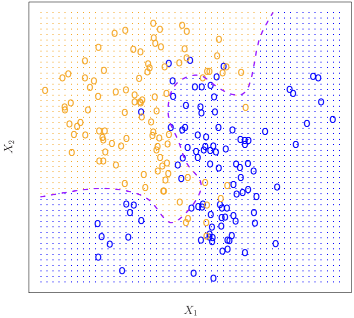
Supervised Learning

- Recall the equation $F = 8.99 \times 10^9 \frac{q_1 q_2}{r^2}$ or more generally $Y = f(\boldsymbol{X})$ where $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$
- In the case above, $Y = F$ and $\boldsymbol{X} = (q_1, q_2, r)^\top$
- Outcome measurement Y (also called dependent variable, response, target)
- Vector of p predictor measurements $\boldsymbol{X}$ (also called inputs, regressors, covariates, features, independent variables)
- In the **regression** problem, Y is quantitative (e.g price, blood pressure)
- In the **classification** problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample)
- We have training data $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_N, Y_N)$. These are observations (examples, instances) of these measurements

  [**Bold** faces correspond to vectors]

Examples of Regression and Classification





Supervised Learning

On the basis of the training data we would like to:

- Accurately predict unseen test cases
- Understand which inputs affect the outcome, and how
- Assess the quality of our predictions and inferences

- Normally from the available examples $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_N, Y_N)$ we choose N_1 of them (around 80%) for training and $N - N_1$ of them for the later evaluation of the model (testing)

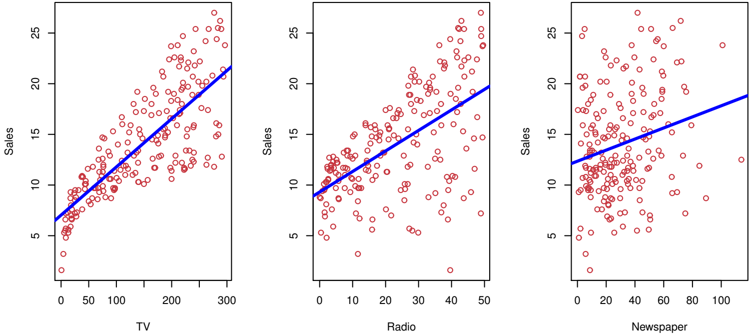
Unsupervised Learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.
- Examples
  - Cocktail Party problem [e.g., see this video]
  - [Just like the Facebook tagging system] you have collection of photos of multiple people, without information who is on which. You want to divide this dataset into multiple piles, each with photos of one individual

More on the Basics of Statistical Learning

An Example

Shown are Sales vs the advertising budget on TV, Radio and Newspaper. Blue linear-regression line fits separately to each:



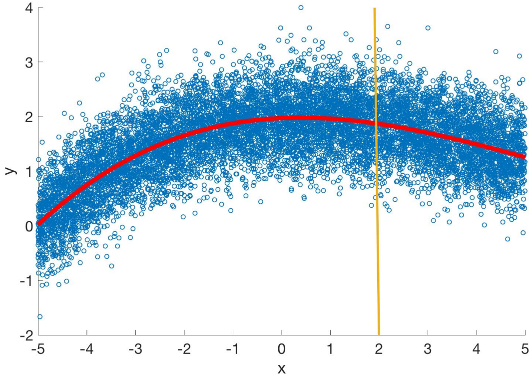
Can we predict Sales using these three? Basically,
$\texttt{sales} = f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$

Why do we Need Predictive Models

- With a good fit we can make predictions of Y at new points $\boldsymbol{X} = \boldsymbol{x}$
- We can understand which features (i.e., components of $\boldsymbol{X} = (X_1, X_2, ..., X_p)^\top$ are important in explaining Y
- For instance in the sales example TV, Radio and Newspaper can be important but "which restaurant is close to the company" might be an irrelevant feature
- Depending on the complexity of the fit, we may be able to understand how each component X_j of $\boldsymbol{X}$ affects Y

Best Predictive Model if We Had Enough Data

- Suppose we had many data samples (X, Y) as below
- What would have been a good estimate of Y for $X = 2$



- In mathematical terms what we are after is $E(Y|X = 2)$

Best Predictive Model if We Had Enough Data

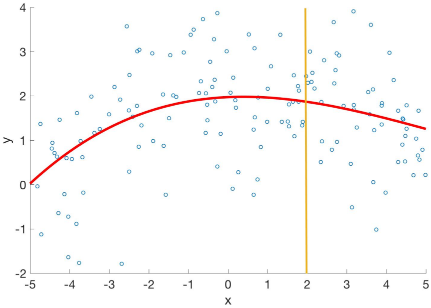
- In an ideal setting of having many data, the reasonable value of Y corresponding to $X = \boldsymbol{x}$ is the average of all responses at $\boldsymbol{X} = \boldsymbol{x}$
- It can be mathematically shown that $f(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$ is the function that minimizes $E[(Y - g(\boldsymbol{X}))^2|\boldsymbol{X} = \boldsymbol{x}]$ over all functions g at all points $\boldsymbol{X} = \boldsymbol{x}$.
- This ideal $f(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$ is called the **regression function**

Question: Can anyone show that $f(x) = E(Y|X = x)$ is the function that minimizes $E[(Y - g(X))^2|X = x]$ over all functions g?

What Happens in Practice

In practice we don't have enough data to have access to f

Even if that could have been the case, doing it for all points $\boldsymbol{x}$ in the space is not practical



One immediate suggestion is considering neighborhoods around $\boldsymbol{x}$

Nearest neighbor averaging can only be good for small p (say < 4) and large N. [Example:] Points $\boldsymbol{x} = (0.9, \ldots, 0.9) \in \mathbb{R}^{100}$ and $\hat{\boldsymbol{x}} = (1, \ldots, 1) \in \mathbb{R}^{100}$ are quite far

Fundamental Limit

- As we observed the best we can do is the regression function

$$Y = f(\boldsymbol{x}) + \epsilon$$

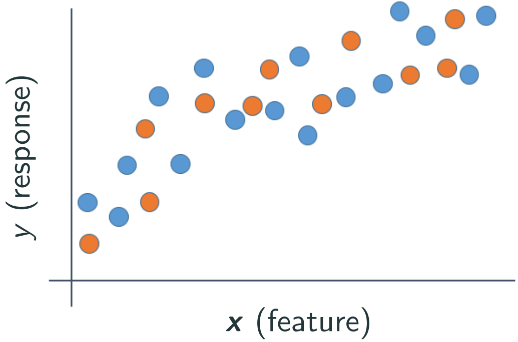
- $f(\boldsymbol{x})$ can be considered as the actual physical model that generates data at $\boldsymbol{X} = \boldsymbol{x}$ and ϵ as the noise and uncertainty
- There is nothing we can do about ϵ, since it is a random variable not under our control
- Even in practice we cannot estimate $f(\boldsymbol{x})$ for $\boldsymbol{X} = \boldsymbol{x}$ and our estimate is something like $\hat{f}(\boldsymbol{x})$ different than $f(\boldsymbol{x})$
- Hence, the overall prediction error at a point $\boldsymbol{X} = \boldsymbol{x}$ is

$$E\left(\left(Y - \hat{f}(\boldsymbol{x})\right)^2 | \boldsymbol{X} = \boldsymbol{x}\right) = \left(f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\right)^2 + \mathrm{Var}(\epsilon)$$

- First term can be improved by improving $\hat{f}(x)$, but the second term is constant and nothing can be done about it

How to Assess Model Accuracy in Practice

- As we said before, we don't normally have access to $f(\boldsymbol{x})$, so how can we assess the accuracy of an estimated model $\hat{f}(\boldsymbol{x})$?
- Usually we split the data into a training set and a test set to later evaluate the accuracy



Parametric vs Nonparametric Models

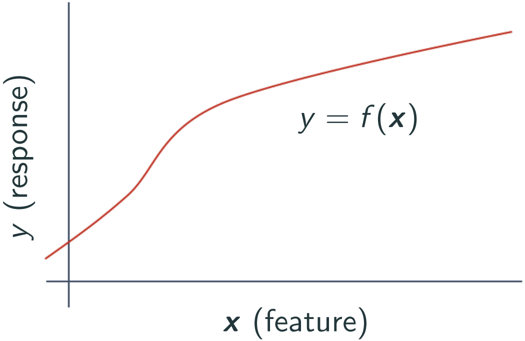
- In parametric models, we use some predefined forms for the model $\hat{f}$ that we would like to fit
- For instance in linear regression we use the following general form and the fitting aims to find the coefficients

$$Y = f_L(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p$$

- Non-parametric methods do not make explicit assumptions about the functional form of the fit. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly
- Thin-plate spline is an example of nonparametric fit, where you can control the smoothness of the fit vs its flexibility to match the training data

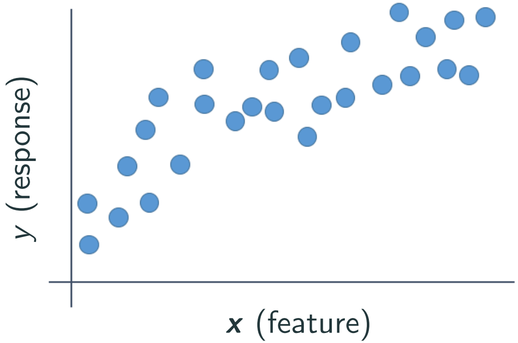
Basics of Model Fitting

- We have a (physical) system which produces a response y to an input $\boldsymbol{x}$



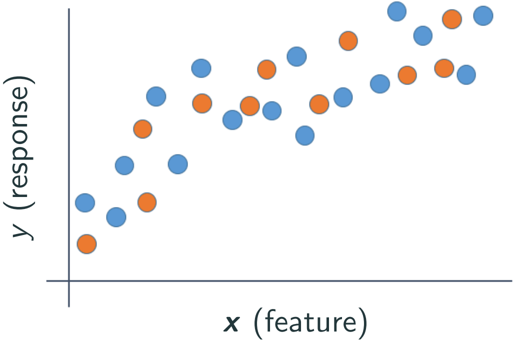
Basics of Model Fitting

- The system is a black box and all we have are samples of input/output



Basics of Model Fitting

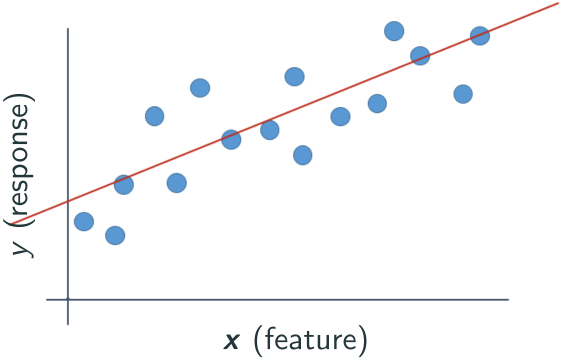
- We may try fitting a model to the data



- We split the data into a training set and a test set to later evaluate the accuracy

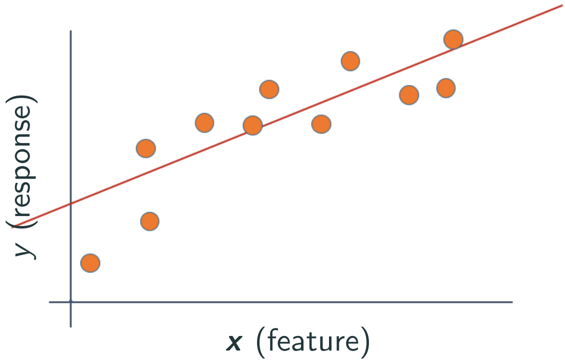
Basics of Model Fitting

- A linear model: $y = w_1 x + w_0$



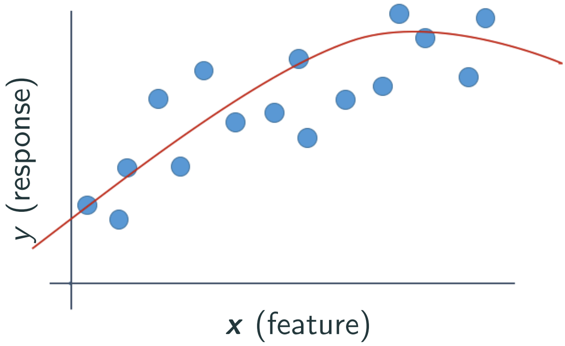
Basics of Model Fitting

- Testing the linear model: $y = \hat{f}(x) = w_1 x + w_0$



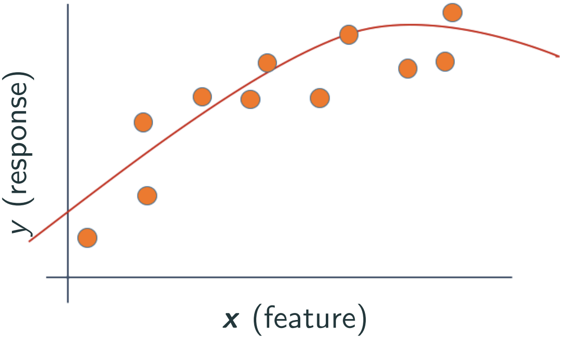
Basics of Model Fitting

- Second order polynomial model: $y = \hat{f}(x) = w_2 x^2 + w_1 x + w_0$



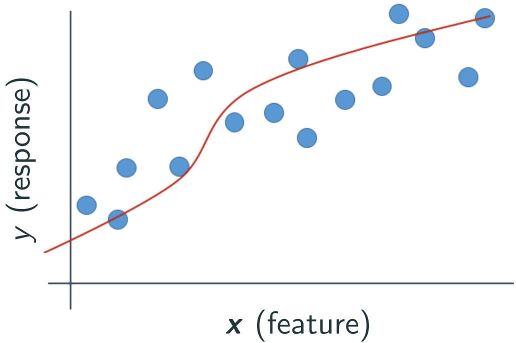
Basics of Model Fitting

- Testing the second order polynomial model:
  $y = \hat{f}(x) = w_2 x^2 + w_1 x + w_0$



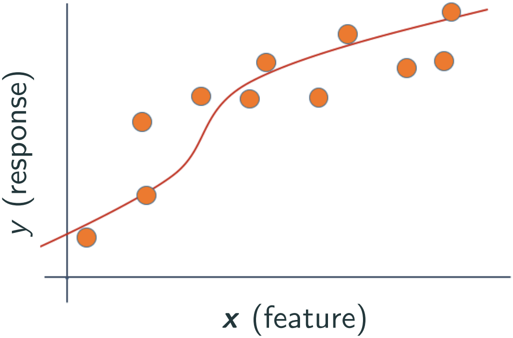
Basics of Model Fitting

- Third order polynomial model: $y = \hat{f}(x) = w_3x^3 + w_2x^2 + w_1x + w_0$



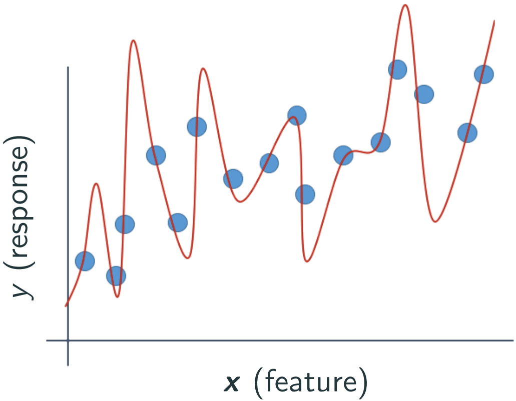
Basics of Model Fitting

- Testing the third order polynomial model:
  $y = \hat{f}(x) = w_3x^3 + w_2x^2 + w_1x + w_0$



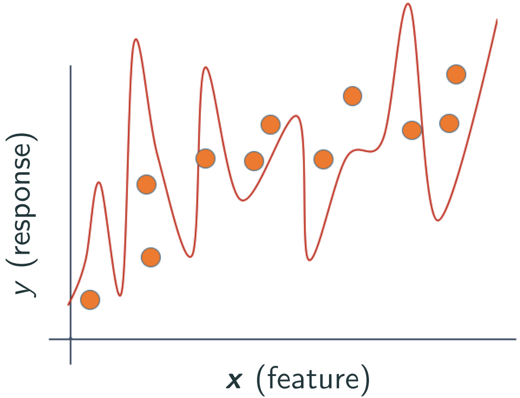
Basics of Model Fitting

- High order polynomial model: $y = \hat{f}(x) = \sum_{p=0}^{P} w_p x^p$ or a thin-plate spline with minor smoothness control



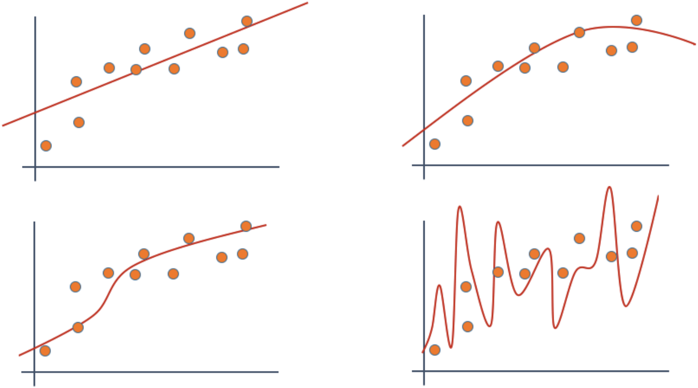
Basics of Model Fitting

- Testing the high order polynomial model: $y = \hat{f}(x) = \sum_{p=0}^{P} w_p x^p$ or a thin-plate spline with minor smoothness control



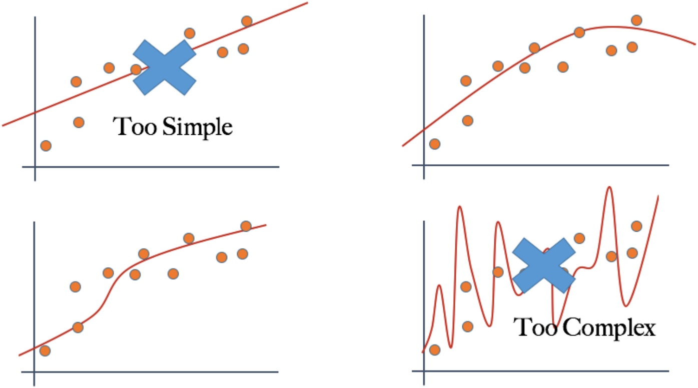
Basics of Model Fitting

- Which model to pick? Underfitting vs overfitting



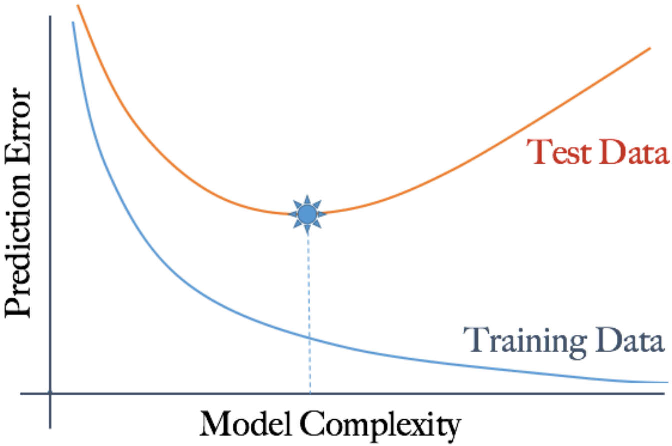
Basics of Model Fitting

- Which model to pick? Underfitting vs overfitting



Typical Curve

- The balance between model flexibility and test error typically looks as below



Assessing Model Accuracy

- Once we fit our model $\hat{f}$ we want to evaluate its performance
- Of course the evaluation should not be with respect to the training data, since that would not be a fair evaluation
- Instead the evaluation is performed on the test data (that has not been used in the fitting)
- Mean Squared Error $= \text{Ave}_{i \in Test}(Y_i - \hat{f}(\boldsymbol{x}_i))^2$

Bias vs Variance(Theory)

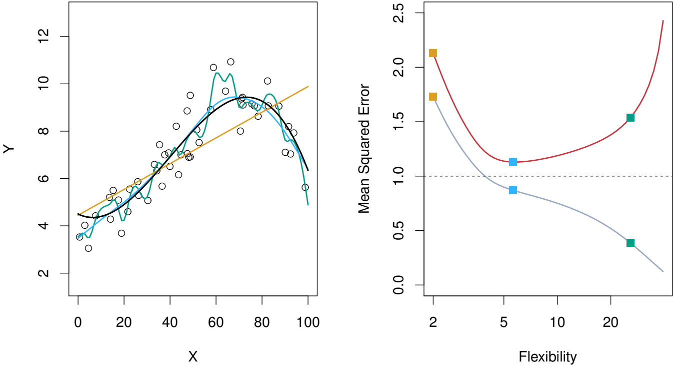
- Suppose we fit a model $\hat{f}$ and we have a large test set to accurately calculate the MSR with respect to the test data
- Assuming that $(\boldsymbol{x}_t, Y_t)$ is a point in the test set and T is the training set, we generally have

$$E_T\left(Y_t - \hat{f}(\boldsymbol{x}_t)\right)^2 = \text{var}(\hat{f}(\boldsymbol{x}_t)) + \left(\text{Bias}(\hat{f}(\boldsymbol{x}))\right)^2 + \text{var}(\epsilon)$$

$$\text{Bias}(\hat{f}(\boldsymbol{x})) = E_T(\hat{f}(\boldsymbol{x}_t) - f(\boldsymbol{x}_t))$$

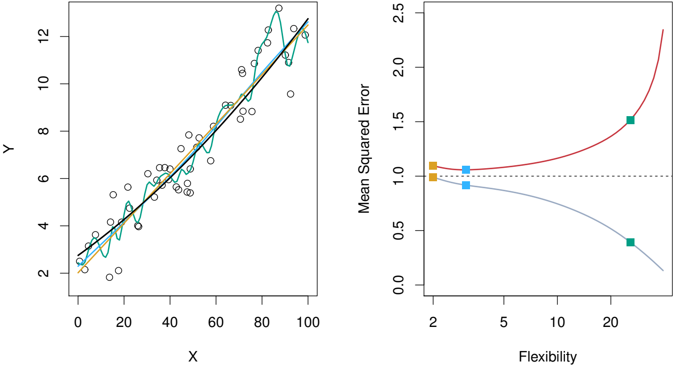
- [Lets derive the equation above]
- Normally as we make $\hat{f}$ more flexible by making it complex (using more sophisticated formulations and more features) the **model variance** term $\text{var}(\hat{f}(\boldsymbol{x}_t))$ increases. This on the other hand decreases the bias (which we want to happen)
- So reducing the test error becomes a trade-off between the bias and the variance

Bias vs Variance



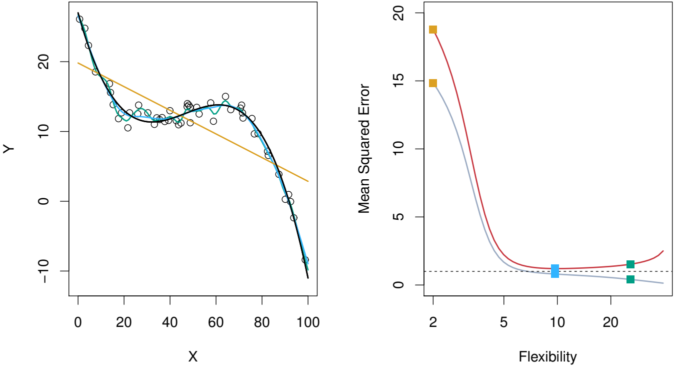
Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

Bias vs Variance



Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

Bias vs Variance



Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

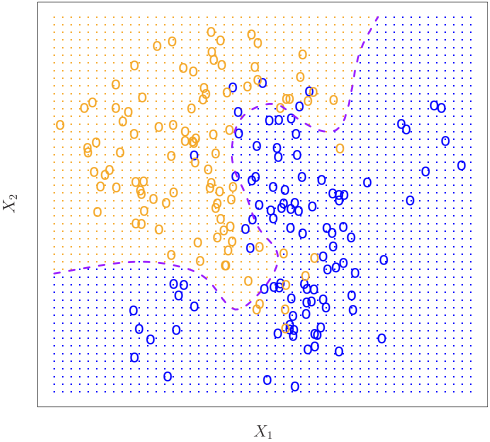
Trade Offs

We just discussed the bias vs variance trade off. In general we deal with various trade offs

- Prediction accuracy versus interpretability:
  e.g., linear models are easy to interpret; thin-plate splines are not.
- Good fit versus over-fit or under-fit:
  How do we know when the fit is just right? (just discussed on the figures above)
- Parsimony versus black-box:
  We often prefer a simpler model involving fewer variables over a black-box predictor involving them all

Classification

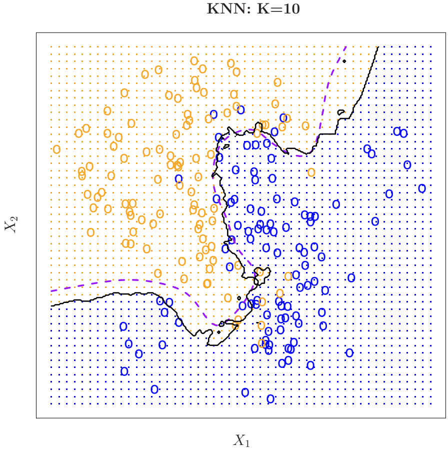
- How overfitting looks for classification problems
- K- Nearest Neighbors Approach



(ideal classifier)

Classification

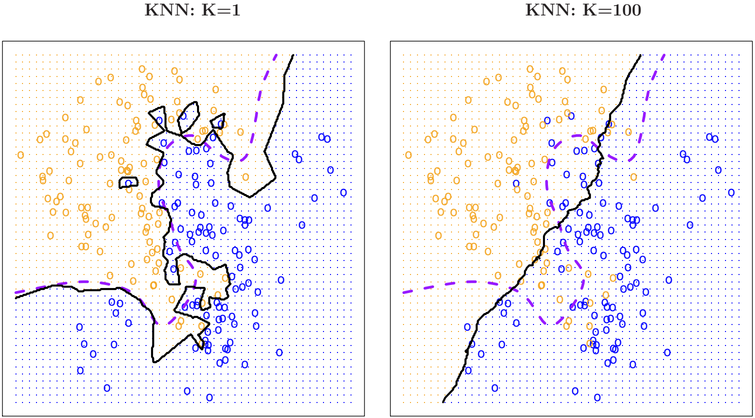
- How overfitting looks for classification problems
- K- Nearest Neighbors Approach (the dense grid is somehow our test)



(KNN Classifier)

Classification

- How overfitting looks for classification problems
- K- Nearest Neighbors Approach (the dense grid is somehow our test)



(KNN Classifier)

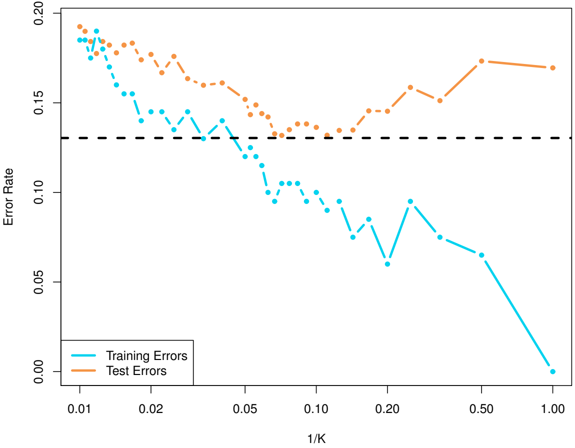
Classification

- K- Nearest Neighbors Approach
- Note that in KNN flexibility is inversely proportional to K

(Just like trying to walk among many people surrounding you)

Classification

- The Trade Off Curve
- K- Nearest Neighbors Approach



(Trade Off)

Lecture Overview

Overview

- Discussed the course information
- Related machine learning to model understanding in science and engineering
- Presented some machine learning examples in real world
- Talked about supervised vs unsupervised learning
- Talked about regression and classification
- Discussed some basics of statistical learning
  - Regression function
  - Test vs training data
  - Overfitting and model trade off
- **Don't worry if some of the topics covered look vague!** We will revisit all this material in depth later in the course

Questions?

References

- https://www.alsharif.info/iom530, 2013.
- J. Friedman, T. Hastie, and R. Tibshirani.
  The elements of statistical learning.
  Springer series in statistics, 2nd edition, 2009.
- G. James, D. Witten, T. Hastie, and R. Tibshirani.
  https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf, 2013.
- G. James, D. Witten, T. Hastie, and R. Tibshirani.
  https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/statistical_learning.pdf, 2013.
- G. James, D. Witten, T. Hastie, and R. Tibshirani.
  An introduction to statistical learning: with applications in R, volume 112.
  Springer, 2013.