Introduction and Logistics

PS690 Computational Methods in Social Science

Jiawei Fu
Department of Political Science
Duke University

Aug 26, 2025

Overview

1. Surprises in high dimensions
2. Statistical Learning
3. Deep learning
4. Topics to be covered
5. Requirements and Grading
6. Bias-Variance Trade-off

High Dimensions

- In the era of "big data," modern social science often works with datasets containing hundreds or even thousands of variables (e.g., administrative records, large-scale surveys).
- In particular, the past few decades have witnessed a dramatic rise in the use of unstructured data such as text, networks, audio, video, and images.
- Classical tools (like OLS with a small set of regressors) break down when $p \gg n$.
- In this course, we will explore cutting-edge methods for high-dimensional data, focusing on prediction, statistical inference, causal inference, and feature extraction.

Surprises in high dimensions

- Our intuition about space is built on two and three dimensions, but it can be highly misleading in high-dimensional settings.
- To build some intuition, let us imagine what it means to buy a "high-dimensional orange."
- Formally, consider a d-dimensional ball (orange) defined as
  $B^d(r) = \{x \in \mathbb{R}^d : \|x\| \leq r\}, \quad \|x\| = \sqrt{x_1^2 + \cdots + x_d^2}.$
- Geometrically, a ball is the region containing all points within a fixed distance r (the radius) from a given point (the center).
- For example:
  - When $d = 2$, the ball is a disk, and its boundary is a circle.
  - When $d = 3$, the ball is the familiar solid sphere.

Surprises in high dimensions

- The volume of a d-dimensional ball is given by

$$V(B^d(r)) = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} r^d,$$

  where $\Gamma(x)$ is the gamma function: $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$, $\Gamma(n+1) = n!$, and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.
- Special cases:
  - When $d = 2$, $V(B^2(r)) = \pi r^2$, the area of a disk.
  - When $d = 3$, $V(B^3(r)) = \frac{4}{3}\pi r^3$, the volume of a solid ball.

Surprises in high dimensions

- First Surprise: The volume of a d-dimensional orange of radius r goes to zero as d increases.

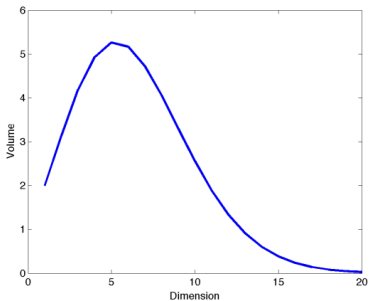


Figure: From [Strohmer, note]

- This means that in high dimensions, your orange is almost entirely empty.

Surprises in high dimensions

- So, where is the pulp of our high-dimensional orange?
- From $\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} r^d$, we note: $V(B^d(r)) = r^d V(B^d(1))$. This shows that the volume of a d-dimensional ball grows exponentially with d (for fixed r).
- Now let us ask: what fraction of the total volume of $B^d(r)$ lies in the thin outer shell of thickness a, i.e., between radius $r - a$ and r?

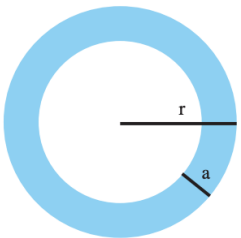
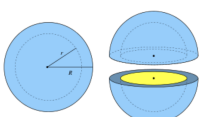


Figure: From [Feres, 2006]

Surprises in high dimensions

- The ratio is
$$\frac{V(B^d(r)) - V(B^d(r-a))}{V(B^d(r))}$$
- Using the fact $V(B^d(r)) = r^d V(B^d(1))$, we get
$$\frac{V(B^d(1))(r^d - (r-a)^d)}{V(B^d(r))r^d} = \frac{r^d - (r-a)^d}{r^d} = 1 - (1 - a/r)^d$$
- Second Surprise: As d increases, almost all of the volume of the ball concentrates near its boundary.
- In other words, if you peel a high-dimensional orange, there is essentially nothing left inside—the pulp is squeezed into an extremely thin layer right next to the peel.

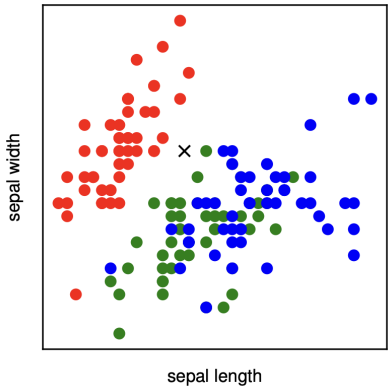


Surprises in high dimensions

- Therefore, in high dimensions, even a dataset that looks "large" is actually very sparse.
- Imagine placing data inside a high-dimensional box: to cover it with small balls of radius r, each ball captures essentially no points. Filling the space requires exponentially many samples.
- This phenomenon is known as the *curse of dimensionality*: many algorithmic approaches in $\mathbb{R}^d$ become exponentially more difficult as the dimension d grows.
- In practice, as dimensionality increases, the number of data points required for reliable performance of any learning algorithm grows exponentially.

A learning example

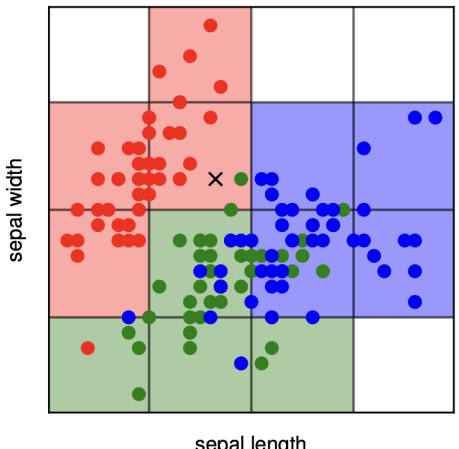
- The figure below shows the Iris dataset, consisting of 50 observations from each of three species of iris flowers.



- Our goal is to predict the species (indicated by color) of the point $\times$.

A learning example

- A very simple algorithm would be to divide the input space into regular cells.
- The identity of the test point is predicted to be the same as the class having the largest number of training points in the same cell as the test point



A learning example

- Let us consider what happens as we increase the number of predictor dimensions.
- If we partition the space into regular cells, the number of cells grows exponentially with the dimension.
- As a result, we would need an exponentially large amount of training data to ensure that most cells are populated rather than empty.

What is statistical learning?

- Suppose we observe a response Y and p predictors $X = (X_1, X_2, \ldots, X_p)$.
- A very general model is

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

  where $f(X) = \mathbb{E}[Y \mid X]$ captures the systematic relationship between X and Y, and ϵ is a random error term.
- In essence, statistical learning consists of a set of methods for estimating the unknown function f.

What is statistical learning?

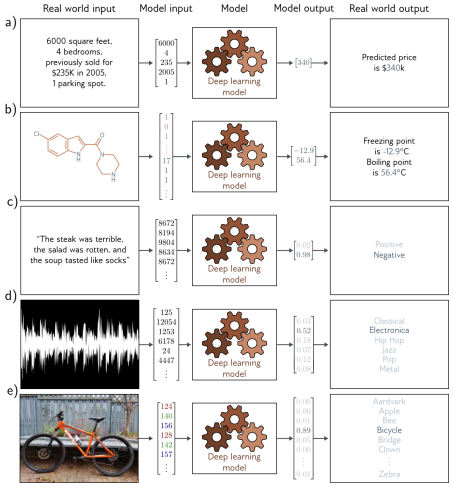
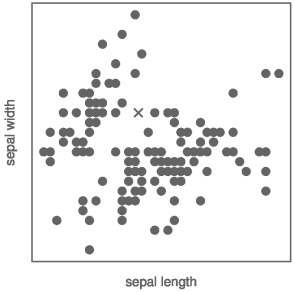


Figure: Supervised learning: from [Prince, 2023]

What is statistical learning?

- The above examples fall under *supervised learning*.
- In contrast, *unsupervised learning* refers to the more challenging setting where we observe X but not the response Y.
- The goal is to uncover or describe the underlying structure of the data.



- For example: What topics do different documents discuss? How can we group or classify documents based on their content?

Deep Learning

- Artificial intelligence (AI) is concerned with building systems that simulate intelligent behavior.
- It encompasses a wide range of approaches, including logic-based reasoning, search algorithms, and probabilistic methods.
- Machine learning (ML) is a subset of AI that makes decisions by fitting mathematical models to data. This field has grown explosively and is now often (though incorrectly) used almost synonymously with AI.
- Deep learning refers to fitting deep neural networks, a particular class of machine learning models.

- Although the curse of dimensionality poses serious challenges for machine learning, it does not prevent us from developing effective methods in high-dimensional spaces.
- One key reason is that real data are often concentrated in a much lower-dimensional subspace, rather than filling the entire ambient space.

  - For example, each image can be represented as a point in a high-dimensional space, where the dimensionality is determined by the number of pixels.
  - However, real images often lie on a much lower-dimensional subspace. In this case, they can be approximately described by just three dimensions: vertical position, horizontal position, and orientation.

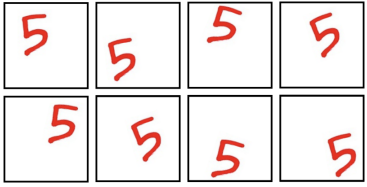


Figure: Caption

- Another way to see that real data are confined to low-dimensional manifolds is to consider the task of generating random images.

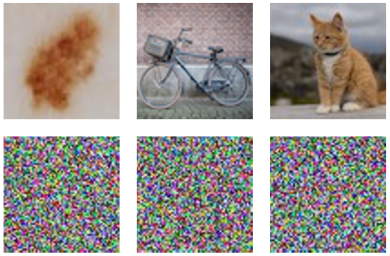


Figure: From [Prince, 2023]

- We observe that none of the synthetic images in the second row resemble natural images.
- In effect, neural networks learn a set of basis functions that are adapted to the underlying data manifold.

Topics to be covered

1. Machine learning: linear model, nonlinear model (tree-based Methods), K-means
2. Statistical Inference after model selection
3. Causal Inference + ML: Double ML and Heterogeneous Treatment Effects
4. Text as Data: Descriptive Inference, feature extraction, Topic Model
5. Causal Inference + Text data
6. Deep learning: CNN, Transformer, NLP and Image, + causal representation?
7. Networks: basic concepts, random-graph model and strategic network formation, applications, and causal inference on networks
8. Your interested topics!

Grading

- (Two) Problem Sets (50%): The most effective way to encourage learning and deepen understanding of the material is through hands-on assignments.
- Although we study machine learning, you also need to learn the machine yourself. Please do not rely entirely on AI to complete your homework.
- AI is most effective when you have built a strong foundation of knowledge yourself. For example, it is well known that generative AI can produce incorrect or misleading information (hallucinations).
- You can only identify these mistakes if you have a solid understanding of the subject.
- You are encouraged to discuss with your peers, but the final work must be written by you.
- For any submitted work, please indicate: which parts, if any, were generated with AI assistance, and who you discussed the assignment with.

Grading

- Midterm (25%): Oct 7. Machine Learning Competition
- I will send out a training dataset for you to learn at home. You may train any model you like.
- On the day of the midterm, I will provide a test dataset in class. You will then use your trained model to make predictions.
- Your grade will be based on the performance of your model.
- One important lesson about machine learning performance is that combining multiple models often leads to better results.
- Since we can only cover a limited set of models in class, you are encouraged to explore additional methods on your own during the weeks without a problem set.

Grading

- Final Project (25%). Due on Dec 15. There are three options.
  1. Apply the Method in Your Own Ongoing Research Project.
  2. Replication and Extension. Select an applied social science paper of interest that uses one of the methods introduced in class or related. Replicate its main findings. In the final section of your replication report, add a new contribution and implement it — such as a meaningful extension, areas for improvement, etc,
  3. Methodological Proposal. If you are more interested in the methodology itself, use this assignment to propose a new research project focusing on the method.

Office Hour

- To save time, please schedule meetings through the appointment on my website (jiaweifu.org) or [link].
- Multiple slots are available throughout the week and will be updated regularly.
- If none of the available times work for you, please send me an email and we will find an alternative time.
- My office is located at 294A Gross Hall.

Resources: Statistics

- Wasserman, L. (2013). All of statistics: a concise course in statistical inference. Springer Science & Business Media.
- Wasserman, L. (2006). All of nonparametric statistics. Springer Science & Business Media.
- Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint (Vol. 48). Cambridge university press.
- Wager, S. (2024, September). Causal inference: A statistical learning approach.
- Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied causal inference powered by ML and AI. arXiv preprint arXiv:2403.02467.

Resources: Statistical Learning

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, No. 1). New York: springer.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
- Fan, J., Li, R., Zhang, C. H., Zou, H. (2020). Statistical foundations of data science. Chapman and Hall/CRC.
- Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.
- Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.
- Bishop, C. M., & Bishop, H. (2023). Deep learning: Foundations and concepts. Springer Nature.
- Prince, S. J. (2023). Understanding deep learning. MIT press.
- Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). Dive into deep learning. Cambridge University Press.

Resources: Text and Networks

- Jackson, M. O. (2008). Social and economic networks. Princeton: Princeton university press.
- Goyal, S., 2023. Networks: An economics approach. MIT Press.
- Jurafsky, D. and Martin, J. H., (2025). Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models.
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as data: A new framework for machine learning and the social sciences. Princeton University Press.

Bias-Variance Trade-off

- One criterion for learning is to minimize difference between true response y_i and predicted one $\hat{y} = \hat{f}(x_i)$.
- Formally, given the data we have, we minimize sample mean squared error (MSE)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- Usually, because we have "seen" current data, and use those data to train the best model we can, this in-sample MSE can be quite small.
- However, does it imply that the model has similar performance on new unseen data?
- There is no guarantee that the method with the lowest training MSE will also have the lowest test MSE.

Training and Test MSE

- In other words, a better criterion is out-of-sample MSE, for a new test point (x_0, y_0):

$$\mathbb{E}[y_0 - \hat{f}(x_0)]^2$$

- Here $\hat{f}$ is random. Sometimes, we can also treat test points as random, can calculate the overall out-of-sample MSE.
- Recall $y = f(x) + \epsilon$, therefore

$$\mathbb{E}[y_0 - \hat{f}(x_0)]^2 = \mathbb{E}[f(x_0) - \hat{f}(x_0)] + Var(\epsilon)$$

- The first part is called *reducible error* because we can potentially improve the accuracy by using the most appropriate statistical learning technique.
- The second part is *irreducible error*, because no matter how well we estimate f, we cannot reduce the error introduced by ϵ.
- We do our best to minimize the reducible error; the best we can achieve is $Var(\epsilon)$.

Training and Test MSE

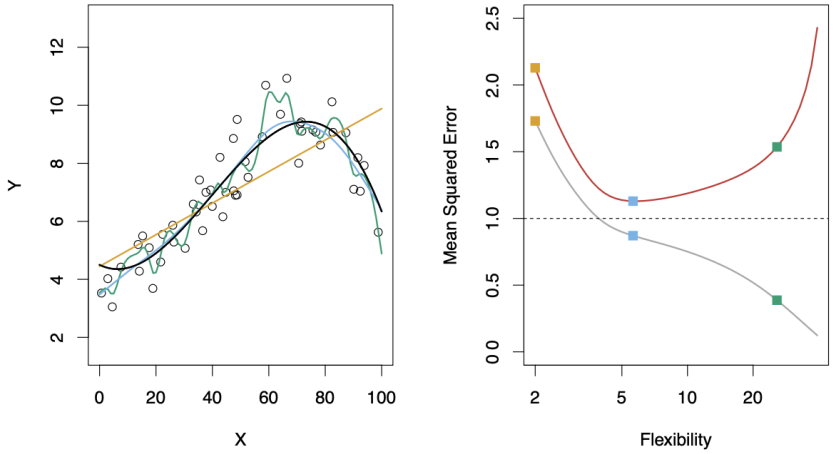


Figure: In-sample (grey curve) and out-of-sample (red) MSE: overfit and underfit

Training and Test MSE

- In practice, we usually could not observe out-of-sample MSE.
- One "default" solution is to separate your data into two parts: training set and test set.
- We learn models in the training set, and then choose the best model based on the test set.
- As we increase the amount of training data, the generalization error typically decreases. Moreover, in general, more data never hurts.

Bias-Variance Trade-off

- Why do we observe a U-shape test MSE?
- Using the trick: $\mathbb{E}[(\theta - \hat{\theta})] = \underbrace{(\theta - \mathbb{E}\hat{\theta})^2}_{Bias^2(\hat{\theta})} + \underbrace{\mathbb{E}[(\hat{\theta} - \hat{\theta})^2]}_{Var(\hat{\theta})}$
- We can further decompose overall out-of-sample MSE (both training and test sample are random):

$$\begin{aligned}\mathbb{E}[y_0 - \hat{f}(x_0)]^2 &= \mathbb{E}[f(x_0) - \hat{f}(x_0)] + Var(\epsilon) \\ &= \mathbb{E}[Bias^2(\hat{f}(x_0))|x_0] + \mathbb{E}[Var(\hat{f}(x_0)|x_0)] + Var(\epsilon)\end{aligned}$$

- We need to select a statistical learning method that simultaneously achieves low variance and low bias.

Bias-Variance Trade-off

- Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
- In general, more flexible statistical methods have higher variance.
- Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
- As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease.

Generalization

- Note that overfitting is not always a bad thing, especially, in deep learning.

Exception (from [Prince, 2023])

At this point, we must stress another important point that we will revisit when introducing deep learning. When a model is capable of fitting arbitrary labels, low training error does not necessarily imply low generalization error. However, it does not necessarily imply high generalization error either! All we can say with confidence is that low training error alone is not enough to certify low generalization error.

Statistical Learning

- For most ML models, we need to choose hyperparameters or say tuning parameters.
- In principle, we should not touch our test set until after we have chosen all our hyperparameters.
- We should never use test data in the model selection process. Otherwise, there is a risk of overfitting.
- If we overfit our training data, there is always the evaluation on test data to keep us honest. But if we overfit the test data, how would we ever know?
- However, we cannot rely solely on the training data for model selection, because we cannot estimate generalization error using the same data on which the model was trained.
- A common solution is to split the data into three parts: a training set, a validation set, and a test set. The validation set is used for model selection, while the test set provides an unbiased evaluation of the final model.

- When training data is scarce, we might not even be able to afford to hold out enough data to constitute a proper validation set. One popular solution to this problem is to employ K-fold cross-validation.
- Here, the original training data is split into non-overlapping subsets. Then model training and validation are executed times, each time training on subsets and validating on a different subset (the one not used for training in that round).
- Finally, the training and validation errors are estimated by averaging over the results from the experiments.

References


Feres, R. (2006).
Lecture note on geometry in very high dimensions.
Lecture note.


Prince, S. J. (2023).
Understanding deep learning.
MIT press.


Strohmer, T. (Lecture note).
Lecture note on surprises in high dimensions.
Lecture note.