CS229 Midterm Review Part II

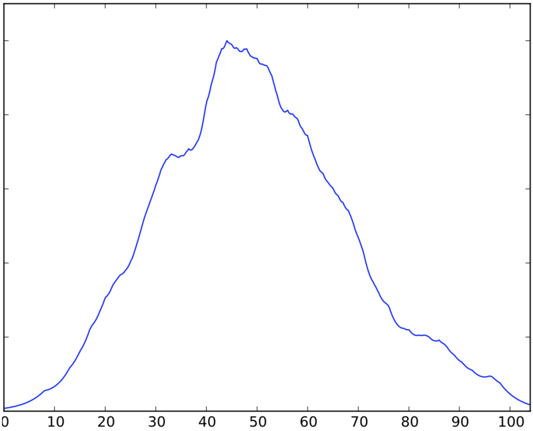
Taide Ding

November 1, 2019

Overview

1. Past Midterm Stats
2. Helpful Resources
3. Notation: quick clarifying review
4. Another perspective on bias-variance
5. Common Problem-solving Strategies (with examples)

The Midterms are tough - DON'T PANIC!

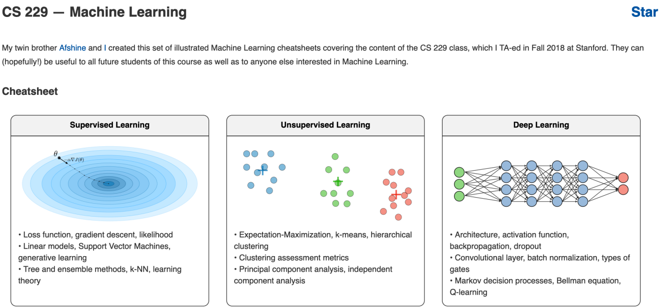


Fall 16 Midterm Grade distribution

Fall 17: $\mu = 39.5, \sigma = 14.5$
Spring 19: $\mu = 65.4, \sigma = 22.4$

Helpful Resources

- Study guide by past CS229 TA Shervine Amidi (link is on course syllabus)



CS 229 — Machine Learning

Star

My twin brother Afshine and I created this set of illustrated Machine Learning cheatsheets covering the content of the CS 229 class, which I TA-ed in Fall 2018 at Stanford. They can (hopefully!) be useful to all future students of this course as well as to anyone else interested in Machine Learning.

Cheatsheet

Supervised Learning	Unsupervised Learning	Deep Learning
• Loss function, gradient descent, likelihood • Linear models, Support Vector Machines, generative learning • Tree and ensemble methods, k-NN, learning theory	• Expectation-Maximization, k-means, hierarchical clustering • Clustering assessment metrics • Principal component analysis, independent component analysis	• Architecture, activation function, backpropagation, dropout • Convolutional layer, batch normalization, types of gates • Markov decision processes, Bellman equation, Q-learning

https://stanford.edu/∼shervine/teaching/cs-229/

Helpful Resources

IMPORTANT: CS229 Linear Algebra and Probability Review handouts

- Go over them carefully and in detail.
- Any and all of the concepts/tools within are fair game w.r.t. solving midterm problems

TAKE NOTES

Notation: quick clarifying review

Notation: quick clarifying review

- $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ denotes a dataset of n examples. For each example i, $x^{(i)} \in \mathbb{R}^d$, and $y^{(i)} \in \mathbb{R}$.

Notation: quick clarifying review

- $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ denotes a dataset of n examples. For each example i, $x^{(i)} \in \mathbb{R}^d$, and $y^{(i)} \in \mathbb{R}$.
- The j-th element (i.e. feature) of the i-th sample is denoted $x_j^{(i)}$.

Notation: quick clarifying review

- $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ denotes a dataset of n examples. For each example i, $x^{(i)} \in \mathbb{R}^d$, and $y^{(i)} \in \mathbb{R}$.
- The j-th element (i.e. feature) of the i-th sample is denoted $x_j^{(i)}$.
- $X \in \mathbb{R}^{n \times d}$ is the data matrix and $\vec{y} \in \mathbb{R}^n$ is the label vector such that:

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix}, X = \begin{bmatrix} - & x^{(1)T} & - \\ - & \vdots & - \\ - & x^{(n)T} & - \end{bmatrix}, \vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}, X\theta = \begin{bmatrix} \theta^T x^{(1)} \\ \vdots \\ \theta^T x^{(n)} \end{bmatrix}$$

for parameter vector $\theta \in \mathbb{R}^d$.

Notation: quick clarifying review

- $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ denotes a dataset of n examples. For each example i, $x^{(i)} \in \mathbb{R}^d$, and $y^{(i)} \in \mathbb{R}$.
- The j-th element (i.e. feature) of the i-th sample is denoted $x_j^{(i)}$.
- $X \in \mathbb{R}^{n \times d}$ is the data matrix and $\vec{y} \in \mathbb{R}^n$ is the label vector such that:

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix}, X = \begin{bmatrix} - & x^{(1)T} & - \\ - & \vdots & - \\ - & x^{(n)T} & - \end{bmatrix}, \vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}, X\theta = \begin{bmatrix} \theta^T x^{(1)} \\ \vdots \\ \theta^T x^{(n)} \end{bmatrix}$$

  for parameter vector $\theta \in \mathbb{R}^d$.
- The t-th iteration of θ is denoted $\theta^{(t)}$.

Notation: quick clarifying review

- $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ denotes a dataset of n examples. For each example i, $x^{(i)} \in \mathbb{R}^d$, and $y^{(i)} \in \mathbb{R}$.
- The j-th element (i.e. feature) of the i-th sample is denoted $x_j^{(i)}$.
- $X \in \mathbb{R}^{n \times d}$ is the data matrix and $\vec{y} \in \mathbb{R}^n$ is the label vector such that:

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix}, X = \begin{bmatrix} - & x^{(1)T} & - \\ - & \vdots & - \\ - & x^{(n)T} & - \end{bmatrix}, \vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}, X\theta = \begin{bmatrix} \theta^T x^{(1)} \\ \vdots \\ \theta^T x^{(n)} \end{bmatrix}$$

  for parameter vector $\theta \in \mathbb{R}^d$.
- The t-th iteration of θ is denoted $\theta^{(t)}$.
- **Superscripts:** sample index $i \in [1, n]$; iteration index $t \in [1, T]$

Notation: quick clarifying review

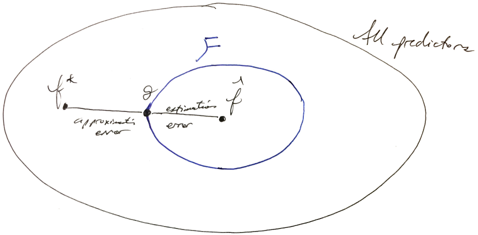
- $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ denotes a dataset of n examples. For each example i, $x^{(i)} \in \mathbb{R}^d$, and $y^{(i)} \in \mathbb{R}$.
- The j-th element (i.e. feature) of the i-th sample is denoted $x_j^{(i)}$.
- $X \in \mathbb{R}^{n \times d}$ is the data matrix and $\vec{y} \in \mathbb{R}^n$ is the label vector such that:

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix}, X = \begin{bmatrix} - & x^{(1)T} & - \\ - & \vdots & - \\ - & x^{(n)T} & - \end{bmatrix}, \vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}, X\theta = \begin{bmatrix} \theta^T x^{(1)} \\ \vdots \\ \theta^T x^{(n)} \end{bmatrix}$$

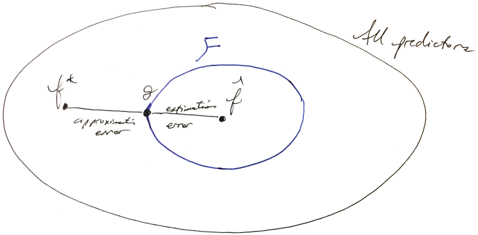
  for parameter vector $\theta \in \mathbb{R}^d$.
- The t-th iteration of θ is denoted $\theta^{(t)}$.
- **Superscripts:** sample index $i \in [1, n]$; iteration index $t \in [1, T]$
- **Subscripts:** feature index $j \in [1, d]$

Another perspective on bias-variance

Another perspective on bias-variance

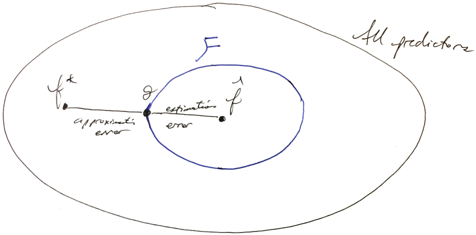


Another perspective on bias-variance



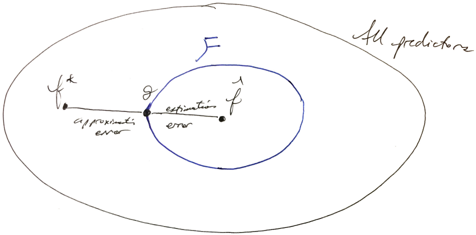
- $\mathcal{F}$ is your model class

Another perspective on bias-variance



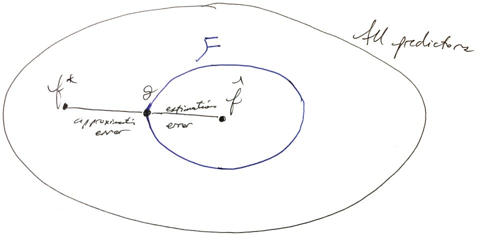
- $\mathcal{F}$ is your model class
- f^* is optimal model for problem (or the true generating distribution)

Another perspective on bias-variance



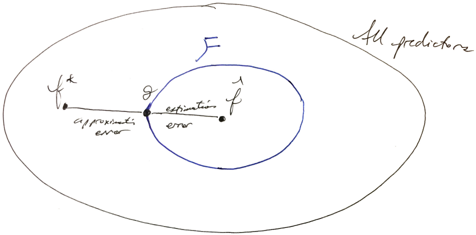
- $\mathcal{F}$ is your model class
- f^* is optimal model for problem (or the true generating distribution)
- g is the optimal model in your model class

Another perspective on bias-variance



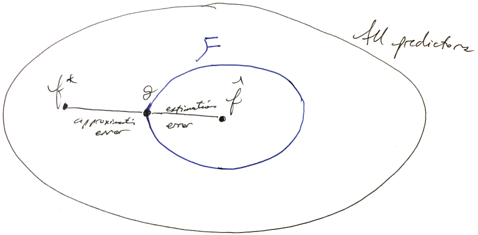
- $\mathcal{F}$ is your model class
- f^* is optimal model for problem (or the true generating distribution)
- g is the optimal model in your model class
- $\hat{f}$ is the model you obtain through learning on your dataset.

Another perspective on bias-variance



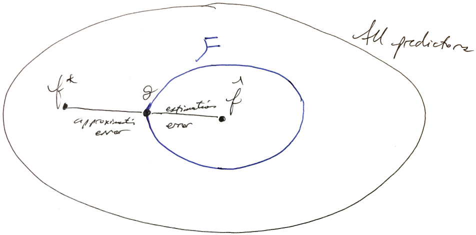
- $\mathcal{F}$ is your model class
- f^* is optimal model for problem (or the true generating distribution)
- g is the optimal model in your model class
- $\hat{f}$ is the model you obtain through learning on your dataset.
- approximation error $\rightarrow$ **bias**

Another perspective on bias-variance



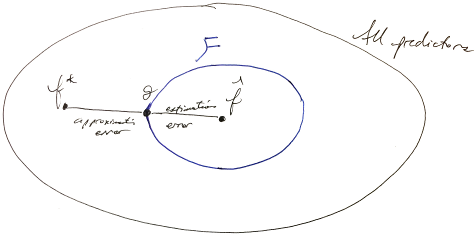
- $\mathcal{F}$ is your model class
- f^* is optimal model for problem (or the true generating distribution)
- g is the optimal model in your model class
- $\hat{f}$ is the model you obtain through learning on your dataset.
- approximation error $\rightarrow$ **bias**
  - reduce bias by expanding $\mathcal{F}$ (e.g. more features, more layers) or moving $\mathcal{F}$ closer to optimal model f^* (i.e. choosing a better class)

Another perspective on bias-variance



- $\mathcal{F}$ is your model class
- f^* is optimal model for problem (or the true generating distribution)
- g is the optimal model in your model class
- $\hat{f}$ is the model you obtain through learning on your dataset.
- approximation error $\rightarrow$ **bias**
  - reduce bias by expanding $\mathcal{F}$ (e.g. more features, more layers) or moving $\mathcal{F}$ closer to optimal model f^* (i.e. choosing a better class)
- estimation error $\rightarrow$ **variance**

Another perspective on bias-variance



- $\mathcal{F}$ is your model class
- f^* is optimal model for problem (or the true generating distribution)
- g is the optimal model in your model class
- $\hat{f}$ is the model you obtain through learning on your dataset.
- approximation error $\rightarrow$ **bias**
  - reduce bias by expanding $\mathcal{F}$ (e.g. more features, more layers) or moving $\mathcal{F}$ closer to optimal model f^* (i.e. choosing a better class)
- estimation error $\rightarrow$ **variance**
  - reduce variance by contracting $\mathcal{F}$ (e.g. remove features, regularize) or making $\vec{f}$ closer to g (e.g. better training algo, more data)

Common Problem-solving Strategies

Common Problem-solving Strategies

Take stock of your arsenal

Common Problem-solving Strategies

Take stock of your arsenal

1. Probability
    - Bayes' Rule
    - Independence, Conditional Independence
    - Chain Rule
    - etc.

Common Problem-solving Strategies

Take stock of your arsenal

1. Probability
   - Bayes' Rule
   - Independence, Conditional Independence
   - Chain Rule
   - etc.
2. Calculus (e.g. taking gradients)
   - Maximum likelihood estimations:

$$\ell(.) = \log \mathcal{L}(.) = \log \prod p(.) = \sum \log p(.)$$

   - Loss minimization
   - etc.

Common Problem-solving Strategies

Take stock of your arsenal

1. Probability
   - Bayes' Rule
   - Independence, Conditional Independence
   - Chain Rule
   - etc.
2. Calculus (e.g. taking gradients)
   - Maximum likelihood estimations:

$$\ell(.) = \log \mathcal{L}(.) = \log \prod p(.) = \sum \log p(.)$$

   - Loss minimization
   - etc.
3. Linear Algebra
   - PSD, eigendecomposition, projection, Mercer's Theorem etc.

Common Problem-solving Strategies

Take stock of your arsenal

1. Probability
   - Bayes' Rule
   - Independence, Conditional Independence
   - Chain Rule
   - etc.
2. Calculus (e.g. taking gradients)
   - Maximum likelihood estimations:

$$\ell(.) = \log \mathcal{L}(.) = \log \prod p(.) = \sum \log p(.)$$

   - Loss minimization
   - etc.
3. Linear Algebra
   - PSD, eigendecomposition, projection, Mercer's Theorem etc.
4. Proof techniques
   - construction, contradiction (e.g. counterexample), induction, contrapositive, etc.

Spring 19 Problem 3(a,b) - Exponential Discr. Analysis

Recall that the Exponential distribution parameterized by $\lambda > 0$ has density

$$p(x; \lambda) = \lambda \exp(-\lambda x), \qquad x \in \mathbb{R}_+.$$

Now suppose that our model is described as follows:

$$\begin{aligned} y &\sim \text{Bernoulli}(\phi) \\ x|y = 0 &\sim \text{Exponential}(\lambda_0) \\ x|y = 1 &\sim \text{Exponential}(\lambda_1) \end{aligned} \tag{2}$$

where ϕ is the parameter of the class marginal distribution, and λ_0 and λ_1 are the class specific parameters for the distribution over input x given $y \in \{0, 1\}$.

(a) [5 points] Derive an exact formula for $p(y = 1|x)$ from the terms defined above, and also show that the resulting classifier has a linear decision boundary in x. Specifically, show that

$$p(y = 1|x) = \frac{1}{1 + \exp\{-(\theta_0 + \theta_1 x)\}}$$

for some θ_0 and θ_1. Clearly state what θ_0 and θ_1 are.

(b) [10 points] Derive the Maximum Likelihood Estimates of ϕ, λ_0 and λ_1 for the given training data using the joint probability (i.e $\ell(\phi, \lambda_0, \lambda_1) = \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \phi, \lambda_0, \lambda_1)$).

Spring 19 Problem 3(a,b) - Exponential Discr. Analysis

Recall that the Exponential distribution parameterized by $\lambda > 0$ has density

$$p(x; \lambda) = \lambda \exp\left(-\lambda x\right), \qquad x \in \mathbb{R}_+.$$

Now suppose that our model is described as follows:

$$\begin{aligned} y &\sim \text{Bernoulli}(\phi) \\ x|y = 0 &\sim \text{Exponential}(\lambda_0) \\ x|y = 1 &\sim \text{Exponential}(\lambda_1) \end{aligned} \tag{2}$$

where ϕ is the parameter of the class marginal distribution, and λ_0 and λ_1 are the class specific parameters for the distribution over input x given $y \in \{0, 1\}$.

(a) [5 points] Derive an exact formula for $p(y = 1|x)$ from the terms defined above, and also show that the resulting classifier has a linear decision boundary in x. Specifically, show that

$$p(y = 1|x) = \frac{1}{1 + \exp\{-(\theta_0 + \theta_1 x)\}}$$

for some θ_0 and θ_1. Clearly state what θ_0 and θ_1 are.

(b) [10 points] Derive the Maximum Likelihood Estimates of ϕ, λ_0 and λ_1 for the given training data using the joint probability (i.e $\ell(\phi, \lambda_0, \lambda_1) = \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \phi, \lambda_0, \lambda_1)$).

Tools used: Probability (Bayes', Indep, Chain Rule), Calculus (MLE)

Spring 19 Problem 5(a) - Kernel Fun

5. **[10 points] Kernel Fun**

   In the following sub-questions, we will explore various properties of Kernels. Throughout the question, we assume $x, z \in \mathbb{R}^d$, $\phi : \mathbb{R}^d \to \mathbb{R}^p$, $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

   (a) [5 points]

   Suppose we have a Positive Semidefinite Matrix $G \in \mathbb{R}^{d \times d}$, and define a function K as follows:

   $$K(x, z) := x^T G z.$$

   Show that K is a valid kernel.

   Remark: Note that G is *not* to be confused to be the kernel matrix.

   Hint: You could consider using eigendecomposition of G, though it is possible to show the result without constructing an explicit feature map.

Tools used: Linear Algebra (PSD properties, eigendecomposition), proof by construction

Summary

1. The midterm is tough. Don't panic!
2. Use resources - study guide, lecture and review handouts, Piazza, OH
3. Know your problem-solving tools - take stock of your arsenal!

Best of Luck!