Statistical Machine Learning: Introduction

Daniel J. McDonald

Indiana University, Bloomington

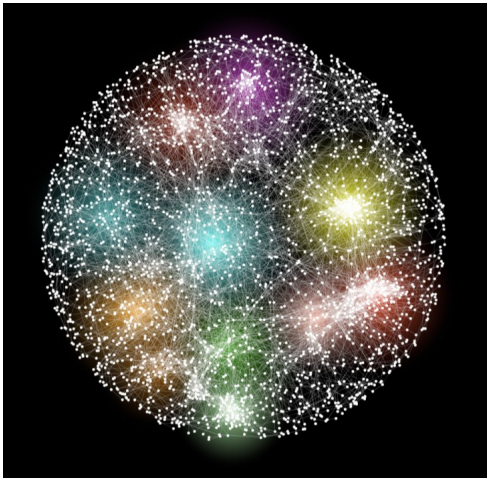
mypage.iu.edu/∼dajmcdon

February 24-26, 2015

Introduction and motivation

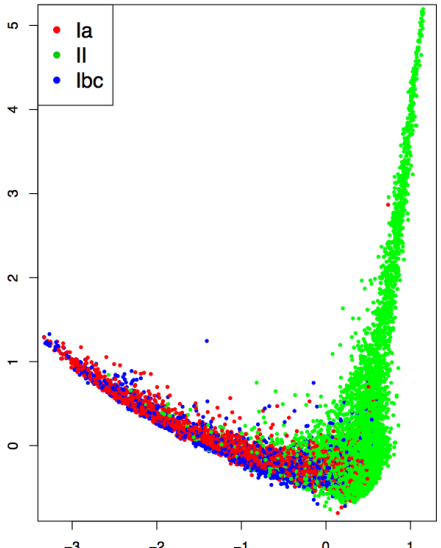
What is statistics?

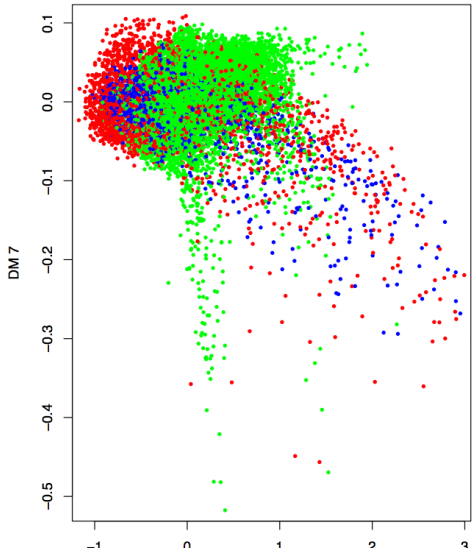
Description Collect some data. Give summaries. Make charts, pretty pictures. Also "unsupervised learning".



What is statistics?

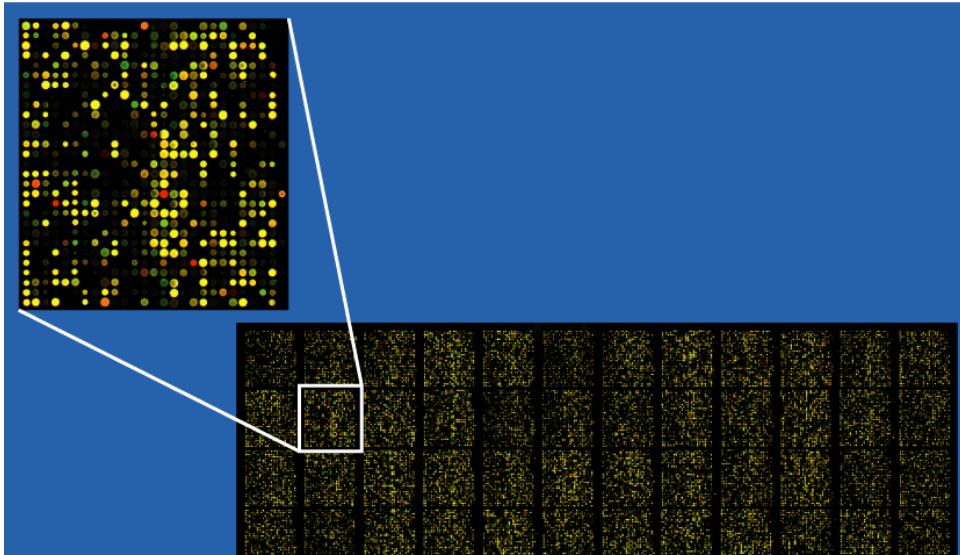
Estimation/Inference Try to determine the underlying causal model.





What is statistics?

Prediction Try to predict some of the data using other data.



What is statistics?

Suppose I come up with some method of description / estimation / prediction. . .

Is it any good?

How can I evaluate my method?

Are there better methods?

Modeling your data

STATISTICAL MODELS

We observe data $Z_1, Z_2, \ldots, Z_n$ generated by some probability distribution P. We want to use the data to learn about P.

A statistical model is a set of distributions $\mathbb{P}$.

Some examples:

1. $\mathbb{P} = \{P(z = 1) = p,\ P(z = 0) = 1 - p,\ 0 < p < 1\}$.
2. $\mathbb{P} = \{Y \sim N(X^\top \beta, \sigma^2),\ \beta \in \mathbb{R}^p,\ \sigma > 0,\ X \text{ fixed}\}$.
3. $\mathbb{P} = \{\text{all CDF's } F\}$.
4. $\mathbb{P} = \{\text{all smooth functions } f : \mathbb{R}^p \to \mathbb{R}\}$

STATISTICAL MODELS

We observe data $Z_1, Z_2, \ldots, Z_n$ generated by some probability distribution P. We want to use the data to learn about P.

$$\mathbb{P} = \{P(z = 1) = p,\ P(z = 0) = 1 - p,\ 0 < p < 1\}$$

To completely characterize P, I just need to estimate p.

Need to assume that $P \in \mathbb{P}$.

This assumption is mostly empty: need independent, can't see $z = 12$.

STATISTICAL MODELS

We observe data $Z_i = (Y_i, X_i)$ generated by some probability distribution P. We want to use the data to learn about P.

$$\mathbb{P} = \{Y \sim N(X^\top \beta, \sigma^2),\ \beta \in \mathbb{R}^p,\ \sigma > 0,\ X \text{ fixed}\}$$

To completely characterize P, I just need to estimate β.

Need to assume that $P \in \mathbb{P}$.

This time, I have to assume a lot more: Linearity, independence, Gaussian noise, no ignored variables, no collinearity, etc

Things you may have seen: Notation

Necessary Background: Notation

- We will write vectors as

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

  We write this as $x \in \mathbb{R}^n$, which is "x is a member of ar-en."

- We commonly will need to "turn" the vector, which we write as

$$x^\top = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

  Here, the superscript "T" takes a vector and flips it on its side.

Necessary Background: Notation

If we have two vectors, we will double subscript them

Suppose $x_1, x_2 \in \mathbb{R}^n$, then

$$x_1 = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{bmatrix} \quad \text{and} \quad x_2 = \begin{bmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2n} \end{bmatrix}$$

Necessary Background: Notation

Often, we will combine many vectors into a matrix

$$\mathbb{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix} = \begin{bmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{bmatrix}$$

As much as possible

- Lower case Roman letters will be columns
- Upper case Roman letters will be rows

Necessary Background: Lengths

We will need to measure the size of both vectors and matrices.

The most common is the one we use every day Euclidean distance

(Think: the Pythagorean theorem)

$$||x||_2 = \sqrt{\sum_{k=1}^{p} x_k^2}$$

We call this a norm and refer to this as the "ell two norm"

Additionally, we will need the Manhattan distance

$$||x||_1 = \sum_{k=1}^{p} |x_k|$$

We call this the "ell one norm"

Necessary Background: Lengths

For matrices, we will just define something very related to 'length'
(but it doesn't technically qualify)

Many times, we are interested in the size of the diagonal of a matrix

This is known as the trace and is defined to be

$$\text{trace}(\mathbb{X}) = \sum_{j=1}^{p} \mathbb{X}_{jj}$$

That is, the trace is the sum of the diagonal entries.

Things you may have seen: Probability

Necessary Background: Probability

For this class, we need to recall (some) probability.

Again, I would be satisfied with you accepting that certain manipulations are reasonable, rather than ‘understanding’ everything.

That being said, let’s take it away...

WHAT'S A RANDOM VARIABLE?

Let X be a random variable. That is, X...

- Has a probability density function p_X such that the probability (denote this by $\mathbb{P}$) that X takes on a set of values A is given by[1]

$$\mathbb{P}(A) = \int_A p_X(x)dx$$

- And p_X has certain properties such as $p_X \geq 0$ and $\int p_X = 1$.

[1] Anyone who has studied probability would have serious problems with this statement. If this is you, don't quibble; we're trying to avoid unnecessary complications.

What are the properties of a random variable?

In this class, we really only care about X's

- mean (alternatively known as its expectation)

  (This is all about finding its center)

- and variance.

  (This is all about finding its spread)

What's expectation?

Imagine taking a metal rod of a certain mass.

However, its mass isn't necessarily even along its length.

Attempt to balance the rod on your finger. The balancing point is the center of mass of the rod.

Figure: A family calculates expectations

WHAT'S EXPECTATION?

Crucial connection: If we think about the density of the random variable determining where the rod's mass is distributed, then the "center of mass" is the expectation.

$$\mathbb{E}[X] = \int x\, p_X(x) dx$$

WHAT'S VARIANCE?

For variance, I'll just give you the definition

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

In words:

"variance is the average squared deviation from the average"

Note: $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

CAN YOU TAKE ME HIGHER?

Implicitly, we were assuming that $X \in \mathbb{R}$.

What happens if $X \in \mathbb{R}^p$?

The expectation is going to look the same, but be a vector

$$\mathbb{E}[X] \in \mathbb{R}^p$$

For variance, we need to use some matrix notation:

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] \in \mathbb{R}^{p \times p}$$

If you write this out, you'll see that this a matrix with

- The variances of the components on the diagonal
- The covariances of any two components in the off-diagonal entries.

COMBINE MATRICES AND PROBABILITY

We will commonly combine matrix multiplication with probability statements

Suppose that $Y \in \mathbb{R}^n$ is a random variable such that $\mathbb{E}[Y] = \mu$ and $\text{Var}[Y] = \Sigma$.

What is the distribution of $\mathbb{X}Y$ (if $\mathbb{X}$ is a fixed matrix, not random)?

It turns out expectation is linear and hence we can rearrange '$\mathbb{E}$' and '$\mathbb{X}$'

$$\mathbb{E}[\mathbb{X}Y] = \mathbb{X}\mathbb{E}[Y] = \mathbb{X}\mu$$

Variance is little more complicated, but not much

$$\text{Var}[\mathbb{X}Y] = \mathbb{X}\,\text{Var}[Y]\,\mathbb{X}^\top = \mathbb{X}\Sigma\mathbb{X}^\top \qquad \text{(check this!)}$$

More on random variables

- If $X_1, \ldots, X_n$ are independent then $\mathrm{Var}[\sum_{i=1}^n a_i X_i] = \sum_i a_i^2 \mathrm{Var}[X_i]$
- The covariance is

$$\mathrm{Cov}[X, Y] = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right] = \mathbb{E}\left[XY\right] - \mu_X \mu_Y$$

- The correlation is

$$\rho(X, Y) = \mathrm{Cov}[X, Y] \,/ \sigma_X \sigma_Y.$$

The conditional expectation of Y given X is the random variable $\mathbb{E}\left[Y|X\right]$.

$$\mathbb{E}\left[Y|X = x\right] = \int y p(y|x) dy$$

IMPORTANT DISTRIBUTIONS

NORMAL $X \sim N(\mu, \sigma^2)$ if

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

BERNOULLI $X \sim Bernoulli(\theta)$ if $P(X=1) = \theta$ and $P(X=0) = 1-\theta$. Thus the pdf is

$$p(x; \theta) = \theta^x (1-\theta)^{1-x} I_{\{0,1\}}(x).$$

CONVERGENCE

Let $X_1, X_2, \ldots$ be a sequence of random variables, and let X be another random variable with distribution P. Let F_n be the cdf of X_n and let F be the cdf of X.

1. X_n converges in probability to X, $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$,

$$\lim_{n \to \infty} P\left(|X_n - X| > \epsilon\right) = 0.$$

2. X_n converges in distribution to X, $X_n \rightsquigarrow X$, if

$$\lim_{n \to \infty} F_n(t) = F(t)$$

Heuristics: the sequence is somehow getting "closer" to some limit. But things are random...

CONVERGENCE RULES

Suppose $X_1, X_2, \ldots$ are independent random variables, each with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

WEAK LAW OF LARGE NUMBERS

$$\overline{X}_n \xrightarrow{P} \mu$$

The law of large numbers tell us that the probability mass of an average of random variables "piles up" near its expectation.

CENTRAL LIMIT THEOREM

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow N(0,1).$$

The CLT tells us about the shape of the "piling", when appropriately normalized.

Who cares?

STATISTICAL MODELS

We observe data $Z_i = (Y_i, X_i)$ generated by some probability distribution P. We want to use the data to learn about P.

$$\mathbb{P} = \{Y \sim N(X^\top \beta, \sigma^2),\ \beta \in \mathbb{R}^p,\ \sigma > 0,\ X \text{ fixed}\}$$

To completely characterize P, I just need to estimate β.

Need to assume that $P \in \mathbb{P}$.

This time, I have to assume a lot more: Linearity, independence, Gaussian noise, no ignored variables, no collinearity, etc

EVALUATION

Once I choose some way to “learn” a statistical model, I need to decide if I’m doing a good job.

How do I decide if I’m doing anything good?

PROPERTIES

Lots of ways to evaluate estimators, $\widehat{\mu}$ of parameters μ.

- Consistency: $\widehat{\mu} \xrightarrow{P} \mu$.
- Asymptotic Normality: $\widehat{\mu} \xrightarrow{D} N(\mu, \Sigma)$
- Efficiency: how large is Σ
- Unbiased: $\mathbb{E}[\widehat{\mu}] \stackrel{?}{=} \mu$
- etc.

None of these things make sense unless your model is correct.

Your model is wrong!

[unless you are flipping coins, gambling in a casino, or running randomized, controlled trials on cereal grains]

MIS-SPECIFIED MODELS

What happens when your model is wrong? And it IS wrong.

None of those evaluation criteria make any sense. The parameters no longer have any meaning.

[The criteria still hold in some sense: I can demand that I get close to the projection of the truth onto $\mathbb{P}$]

PREDICTION

Prediction is easier: your model may not actually represent the true state of nature, but it may still predict well.

Over an 13-year period, [David Leinweber] found, [that annual butter production in Bangladesh] "explained" 75% of the variation in the annual returns of the Standard & Poor's 500-stock index.

By tossing in U.S. cheese production and the total population of sheep in both Bangladesh and the U.S., Mr. Leinweber was able to "predict" past U.S. stock returns with 99% accuracy.

via Carl Richards, NYT 3/26/2012

The predictive viewpoint

THE SETUP

What do we mean by good predictions?

We make observations and then attempt to "predict" new, unobserved data.

Sometimes this is the same as estimating the mean.

Mostly, we observe $(y_1, x_1), \ldots, (y_n, x_n)$, and we want some way to predict Y from X.

EVALUATING PREDICTIONS

Of course, both Y and $\widehat{Y}$ are random

I want to know how well I can predict on average

Let $\widehat{f}$ be some way of making predictions $\widehat{Y}$ of Y using covariates X

In fact, suppose I observe a dataset $\mathcal{D}_n = \{(Y_1, X_1,), \ldots, (Y_n, X_n)\}$. Then I want to choose some $\widehat{f}$ using $\mathcal{D}_n$.

Is $\widehat{f}$ good on average?

EVALUATING PREDICTIONS

Choose some loss function that measures prediction quality: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. We predict Y with $\widehat{Y}$

Examples:

$$\begin{aligned}
&\text{SQUARED-ERROR:} && \ell(y, \widehat{y}) = (y - \widehat{y})^2 \\
&\text{ABSOLUTE-ERROR:} && \ell(y, \widehat{y}) = |y - \widehat{y}| \\
&\text{ZERO-ONE:} && \ell(y, \widehat{y}) = I(y \neq \widehat{y}) = \begin{cases} 0 & y = \widehat{y} \\ 1 & \text{else} \end{cases}
\end{aligned}$$

Can be generalized to Y in arbitrary spaces.

Prediction risk

Prediction risk

$$R_n(\widehat{f}) = \mathbb{E}[\ell(Y, \widehat{f}(X))]$$

where the expectation is taken over the new data point (Y, X) and $\mathcal{D}_n$ (everything that is random).

For regression applications, we will use squared-error loss:

$$R_n(\widehat{f}) = \mathbb{E}[(Y - \widehat{f}(X))^2]$$

For classification applications, we will use zero-one loss:

$$R_n(\widehat{f}) = \mathbb{E}[I(Y \neq \widehat{f}(X))]$$

Example 1: The mean

ESTIMATING THE MEAN

Suppose we know that we want to predict a quantity Y, where $\mathbb{E}[Y] = \mu \in \mathbb{R}$ and $\mathrm{Var}[Y] = 1$. That is, $Y \sim P \in \mathbb{P}$, where

$$\mathbb{P} = \{P : \mathbb{E}[Y] = \mu \text{ and } \mathrm{Var}[Y] = 1\}.$$

Our data is $\mathcal{D}_n = \{Y_1, \ldots, Y_n\}$ such that $Y_i \overset{i.i.d.}{\sim} P$, and we want to estimate μ (and hence P).

ESTIMATING THE MEAN

Let $\widehat{Y} = \overline{Y}_n$ be the sample mean.

We can ask about the estimation risk (since we're estimating μ):

$$\begin{aligned} R_n(\overline{Y}_n; \mu) &= \mathbb{E}[(\overline{Y}_n - \mu)^2] \\ &= \mathbb{E}[\overline{Y}_n^2] - 2\mu\mathbb{E}[\overline{Y}_n] + \mu^2 \\ &= \mu^2 + \frac{1}{n} - 2\mu^2 + \mu^2 \\ &= \frac{1}{n} \end{aligned}$$

PREDICTING NEW Y'S

Let $\widehat{Y} = \overline{Y}_n$ be the sample mean. This is a better idea.

What is the prediction risk of $\overline{Y}$?

$$\begin{aligned} R_n(\overline{Y}_n) &= \mathbb{E}[(\overline{Y}_n - Y)^2] \\ &= \mathbb{E}[\overline{Y}_n^2] - 2\mathbb{E}[\overline{Y}_n Y] + \mathbb{E}[Y^2] \\ &= \mu^2 + \frac{1}{n} - 2\mu^2 + \mu^2 + 1 \\ &= 1 + \frac{1}{n} \end{aligned}$$

PREDICTING NEW Y'S

What is the prediction risk of guessing $Y = 0$?

You can probably guess that this is a stupid idea.

Let's show why it's stupid.

$$\begin{aligned} R_n(0) &= \mathbb{E}[(0 - Y)^2] \\ &= \mathbb{E}[(Y - \mu + \mu)^2] \\ &= \mathbb{E}[(Y - \mu)^2] + 2\mu\mathbb{E}[(Y - \mu)] + \mu^2 \\ &= 1 + \mu^2 \end{aligned}$$

PREDICTING NEW Y'S

What is the prediction risk of guessing $Y = \mu$?

This is a great idea, but we don't know μ.

Let's see what happens anyway.

$$\begin{aligned} R_n(\mu) &= \mathbb{E}[(Y - \mu)^2] \\ &= 1 \end{aligned}$$

ESTIMATING THE MEAN

Prediction risk: $R(\overline{Y}_n) = 1 + \frac{1}{n} = \sigma^2 + \frac{\sigma^2}{n}$
Estimation risk: $R(\overline{Y}_n; \mu) = \frac{1}{n}$

There is actually a nice interpretation here:

The common $1/n$ term is $\text{Var}\left[\overline{Y}_n\right]$

The extra factor of 1 in the prediction risk is irreducible error — Y is a random variable, and hence noisy. We can never eliminate it's intrinsic variance. In other words, even if we knew μ, we could never get closer than 1, on average.

Intuitively, $\overline{Y}_n$ is the obvious thing to do.

Predicting new Y's

Let's try one more: $\widehat{Y}_a = a\overline{Y}_n$ for some $a \in (0, 1]$.

$$R_n(\widehat{Y}_a) = \mathbb{E}[(\widehat{Y}_a - Y)^2] = (1-a)^2\mu^2 + \frac{a^2}{n} + 1$$

We can minimize this in a to get the best possible prediction risk for an estimator of the form $\widehat{Y}_a$:

$$\operatorname*{argmin}_a R_n(\widehat{Y}_a) = \left(\frac{\mu^2}{\mu^2 + 1/n}\right)\overline{Y}_n$$

What happens if $\mu \ll 1$?

Wait a minute! You're saying there is a better estimator than $\overline{Y}_n$?

PREDICTING NEW Y'S

Let's try one more: $\widehat{Y}_a = a\overline{Y}_n$ for some $a \in (0, 1]$.

$$R_n(\widehat{Y}_a) = \mathbb{E}[(\widehat{Y}_a - Y)^2] = (1-a)^2\mu^2 + \frac{a^2}{n} + 1$$

We can minimize this in a to get the best possible prediction risk for an estimator of the form $\widehat{Y}_a$:

$$\operatorname*{argmin}_a R_n(\widehat{Y}_a) = \left(\frac{\mu^2}{\mu^2 + 1/n}\right)\overline{Y}_n$$

What happens if $\mu \ll 1$?

Wait a minute! You're saying there is a better estimator than $\overline{Y}_n$?

Why deal with prediction risk?

PREDICTION RISK

$$R_n(f) = \mathbb{E}[\ell(Y, f(X))]$$

Why care about $R_n(f)$?

Measures predictive accuracy on average.

How much confidence should you have in f's predictions.

Compare with other models.

This is hard:
Don't know P (if I knew the truth, this would be easy)

What if you really want to make inferences?

1. You don't really care about predicting what will happen next year / quarter / millisecond
2. But you do want to offer an explanation / evaluate counterfactuals / describe the world
3. So you need the structure of your model to be at least approximately right
4. If you cannot predict well, then your model cannot be correct at all
5. Therefore, a necessary condition to believe your counterfactuals is good predictive accuracy

- Step 4 is about prediction error. (Sum of squared residuals??)
- Best not be fooling yourself in step 5.

RISK FOR GENERAL MODELS

We just saw that when you know the true model, and you have a nice estimator, the prediction risk has a nice decomposition

(this generalizes to much more complicated situations)

- Suppose we have a class of prediction functions $\mathcal{F}$,

$$\text{e.g. } \mathcal{F} = \left\{\beta : f(x) = \beta^\top x\right\}$$

- We use the data to choose some $\widehat{f} \in \mathcal{F}$ and set $\widehat{Y} = \widehat{f}(X)$
- The true model is g (not necessarily in $\mathcal{F}$). Then:

$$R_n(\widehat{f}) = \int \left[\text{bias}^2(\widehat{f}(x)) + \text{var}(\widehat{f}(x))\right] p(x)dx + \sigma^2$$

where $X \sim p$ and

$$\begin{aligned}
\text{bias}(\widehat{f}(x)) &= \mathbb{E}[\widehat{f}(x)] - g(x) \\
\text{var}(\widehat{f}(x)) &= \mathbb{E}[(\widehat{f}(x) - \mathbb{E}\widehat{f}(x))^2] \\
\sigma^2 &= \mathbb{E}[(Y - g(X))^2]
\end{aligned}$$

BIAS-VARIANCE DECOMPOSITION

So,

$$\begin{array}{lcllll} \text{prediction risk} & = & \text{bias}^2 & + & \text{variance} & + & \text{irreducible error} \\ \text{estimation risk} & = & \text{bias}^2 & + & \text{variance} & & \end{array}$$

What is $R(a)$ for our estimator $\widehat{Y}_a = a\overline{Y}_n$?

$$\text{bias}(\widehat{Y}_a) = \mathbb{E}[a\overline{Y}_n] - \mu = (a-1)\mu$$

$$\text{var}(\widehat{f}(x)) = \mathbb{E}[(a\overline{Y}_n - \mathbb{E}[a\overline{Y}_n])^2] = a^2\mathbb{E}[(\overline{Y}_n - \mu)^2] = \frac{a^2}{n}$$

$$\sigma^2 = \mathbb{E}[(Y-\mu)^2] = 1$$

$$\left(\text{That is: } R_n(\widehat{Y}_a) = (a-1)^2\mu^2 + \frac{a^2}{n} + 1\right)$$

BIAS-VARIANCE DECOMPOSITION

So,

$$\begin{array}{lcllcl} \text{prediction risk} & = & \text{bias}^2 & + & \text{variance} & + \quad \text{irreducible error} \\ \text{estimation risk} & = & \text{bias}^2 & + & \text{variance} & \end{array}$$

What is $R(a)$ for our estimator $\widehat{Y}_a = a\overline{Y}_n$?

$$\text{bias}(\widehat{Y}_a) = \mathbb{E}[a\overline{Y}_n] - \mu = (a-1)\mu$$

$$\text{var}(\widehat{f}(x)) = \mathbb{E}[(a\overline{Y}_n - \mathbb{E}[a\overline{Y}_n])^2] = a^2\mathbb{E}[(\overline{Y}_n - \mu)^2] = \frac{a^2}{n}$$

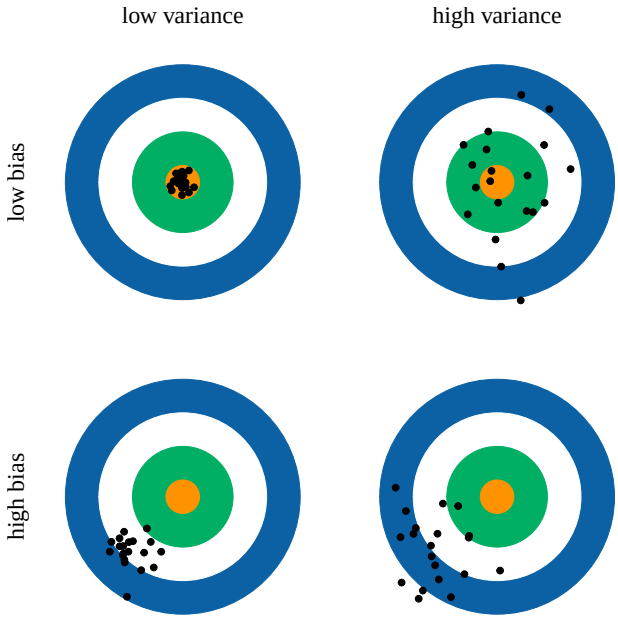
$$\sigma^2 = \mathbb{E}[(Y-\mu)^2] = 1$$

$$\left(\text{That is: } R_n(\widehat{Y}_a) = (a-1)^2\mu^2 + \frac{a^2}{n} + 1\right)$$

BIAS-VARIANCE DECOMPOSITION

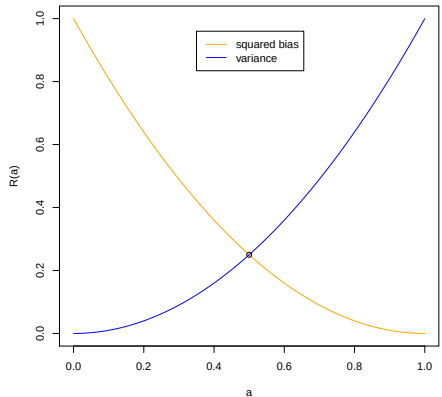
Important implication: prediction risk is proportional to estimation risk. However, defining estimation risk requires stronger assumptions.

In order to make good predictions, we want our prediction risk to be small. This means that we want to 'balance' the bias and variance.


low variance
high variance
low bias
high bias

BIAS-VARIANCE TRADEOFF: ESTIMATING THE MEAN

$$R(a) = R_n(\widehat{Y}_a) = (a-1)^2\mu^2 + \frac{a^2}{n} + 1$$



Here, $\mu = 1 (= \sigma^2)$ and $n = 1$, and hence optimal $a = .5$. This corresponds to the intersection of these curves.

WHAT?

Just to restate:

If $\mu = 1$ and $n = 1$ then it is better to predict with $.5Y$ than with Y itself.

In this case

$$R(a) = R_1(aY) = 1.5$$
$$R(Y) = 1 + 1 = 2$$

BIAS-VARIANCE TRADEOFF: OVERVIEW

- bias: how well does $\widehat{f}$ approximate the truth g
- more complicated $\mathcal{F}$, lower bias. Flexibility $\Rightarrow$ Parsimony
- more flexibility $\Rightarrow$ larger variance
- complicated models are hard to estimate precisely for fixed n
- irreducible error

Example 2: Normal means

NORMAL MEANS

Suppose we observe the following data:

$$Y_i = \beta_i + \epsilon_i, \qquad i = 1, \ldots, n$$

where $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, 1)$.

We want to estimate $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$.

The maximum likelihood estimator is $\widehat{\boldsymbol{\beta}}^{MLE} = (Y_1, \ldots, Y_n)$.

This estimator has lots of nice properties: consistent, unbiased, UMVUE, (asymptotic) normality

But the MLE STINKS! It's a bad estimator.

It has no bias, but big variance.

$$R_n(\widehat{\boldsymbol{\beta}}^{MLE}) = \text{bias}^2 + \text{var} = 0 + n \cdot 1 = n$$

What if we use a biased estimator?

Consider the following estimator instead:

$$\widehat{\beta}_i^S = \begin{cases} Y_i & i \in S \\ 0 & \text{else.} \end{cases}$$

Here $S \subseteq \{1, \ldots, n\}$.

NORMAL MEANS

$$R_n(\widehat{\boldsymbol{\beta}}^S) = \sum_{i \notin S} \beta_i^2 + |S|.$$

In other words, if some $\beta_i < 1$, then don't bother estimating them!

In general, introduced parameters like S will be called tuning parameters.

Of course we don't know which $\beta_i < 1$.

But we could try to estimate $R_n(\widehat{\boldsymbol{\beta}}^S)$, and choose S to minimize our estimate.

ESTIMATING R_n

By definition, for any estimator $\widehat{\boldsymbol{\beta}}$,

$$R_n(\widehat{\boldsymbol{\beta}}) = \mathbb{E}\left[\sum_{i=1}^{n}(\widehat{\beta}_i - \beta_i)^2\right]$$

An intuitive estimator of R_n is

$$\widehat{R}_n(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(\widehat{\beta}_i - Y_i)^2.$$

This is known as the training error and it can be shown that

$$\widehat{R}_n(\widehat{\boldsymbol{\beta}}) \approx R_n(\widehat{\boldsymbol{\beta}}).$$

Also,

$$\widehat{\boldsymbol{\beta}}^{MLE} = \underset{\beta}{\operatorname{argmin}}\, \widehat{R}_n(\widehat{\boldsymbol{\beta}}^{MLE}).$$

What could possibly go wrong?

DANGERS OF USING THE TRAINING ERROR

Although

$$\widehat{R}_n(\widehat{\boldsymbol{\beta}}) \approx R_n(\widehat{\boldsymbol{\beta}}),$$

this approximation can be very bad. In fact:

$$\begin{array}{lrcl} \text{TRAINING ERROR:} & \widehat{R}_n(\widehat{\boldsymbol{\beta}}^{MLE}) & = & 0 \\ \text{RISK:} & R_n(\widehat{\boldsymbol{\beta}}^{MLE}) & = & n \end{array}$$

In this case, the optimism of the training error is n.

What about $\widehat{\boldsymbol{\beta}}^S$?

$$\widehat{R}_n(\widehat{\boldsymbol{\beta}}^S) = \sum_{i=1}^{n}(\widehat{\beta}_i - Y_i)^2 = \sum_{i \notin S} Y_i^2$$

Well

$$\mathbb{E}\left[\widehat{R}_n(\widehat{\boldsymbol{\beta}}^S)\right] = R_n(\widehat{\boldsymbol{\beta}}^S) - 2|S| + n.$$

So I can choose S by minimizing $\widehat{R}_n(\widehat{\boldsymbol{\beta}}^S) + 2|S|$.

$$\text{Estimate of Risk} = \text{training error} + \text{penalty}.$$

The penalty term corrects for the optimism.

Where we’re going

THEMES OF COURSE

BIAS IS GOOD

1. Very often, we can trade some bias for (much) lower variance
2. Bias is controlled by setting tuning parameters (e.g. S)
3. Choosing tuning parameters carefully gives good risk properties
4. To know how to choose the tuning parameters, we need an estimate of the risk
5. Training error (which a risk estimator) is a bad choice (optimistic)
6. Unbiased estimators of parameters in correct models may have nice properties
7. Unbiased estimators of parameters in mis-specified models rarely have nice properties
8. All models are mis-specified

Up next...

1. Regression and regularization
2. Economic forecasting and time series
3. Classification
4. More fun stuff...