Learning Theory, Overfitting, Bias Variance Decomposition

Machine Learning 10-601B

Seyoung Kim

Many of these slides are derived from Tom Mitchell, Ziv-Bar Joseph. Thanks!

How many examples will ϵ-exhaust the VS?

Theorem: [Haussler, 1988].

If the hypothesis space H is finite, and D is a sequence of $m \geq 1$ independent random examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that the version space with respect to H and D is not ϵ-exhausted (with respect to c) is less than

$$|H|e^{-\epsilon m}$$

Interesting! This bounds the probability that <u>any consistent learner</u> will output a hypothesis h with $error(h) \geq \epsilon$

<u>Any(!)</u> learner that outputs a hypothesis consistent with all training examples (i.e., an h contained in $VS_{H,D}$)

What it means

[Haussler, 1988]: probability that the version space is not ε-exhausted after *m* training examples is at most $|H|e^{-\epsilon m}$

$\uparrow$

Suppose we want this probability to be at most δ

$$\Pr[(\exists h \in H) s.t. (error_{train}(h) = 0) \wedge (error_{true}(h) > \epsilon)] \leq |H|e^{-\epsilon m}$$

1. How many training examples suffice?

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Agnostic Learning

So far, assumed $c \in H$

Agnostic learning setting: don't assume $c \in H$

- What do we want then?
  - The hypothesis h that makes fewest errors on training data
- What is sample complexity in this case?

$$m \geq \frac{1}{2\epsilon^2}(\ln |H| + \ln(1/\delta))$$

Here ε is the difference between the training error and true error of the output hypothesis (the one with lowest training error)

Additive Hoeffding Bounds – Agnostic Learning

- Given *m* independent flips of a coin with true Pr(heads) = θ
  we can bound the error ϵ in the maximum likelihood estimate $\widehat{\theta}$

$$\Pr[\theta > \hat{\theta} + \epsilon] \leq e^{-2m\epsilon^2}$$

- Relevance to agnostic learning: for any *single* hypothesis *h*

$$\Pr[error_{true}(h) > error_{train}(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

- But we must consider all hypotheses in H

$$\Pr[(\exists h \in H) error_{true}(h) > error_{train}(h) + \epsilon] \leq |H| e^{-2m\epsilon^2}$$

- Now we assume this probability is bounded by δ. Then, we have

$$m > \frac{1}{\varepsilon^2}(\ln |H| + \ln(1/\delta))$$

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Question: If H = {h | h: X → Y} is infinite, what measure of complexity should we use in place of |H| ?

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Question: If H = {h | h: X → Y} is infinite, what measure of complexity should we use in place of |H| ?

Answer: The largest subset of X for which H can <u>guarantee</u> zero training error (regardless of the target function c)

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

Question: If H = {h | h: X → Y} is infinite, what measure of complexity should we use in place of |H| ?

Answer: The largest subset of X for which H can guarantee zero training error (regardless of the target function c)

VC dimension of H is the size of this subset

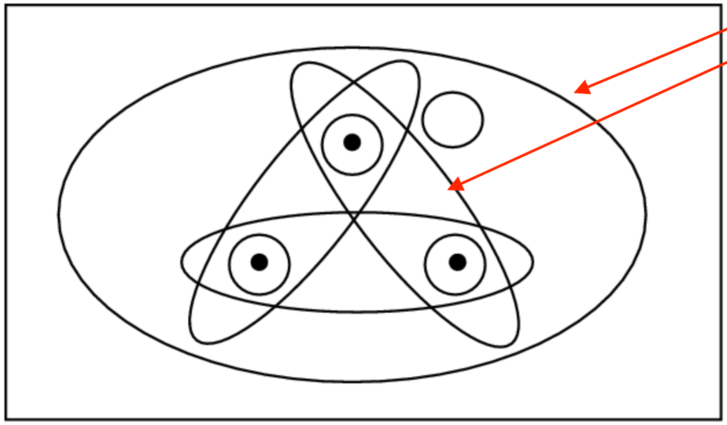
Shattering a Set of Instances

Definition: a **dichotomy** of a set S is a partition of S into two disjoint subsets.

a labeling of each member of S as positive or negative

Definition: a set of instances S is **shattered** by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

Instance space X



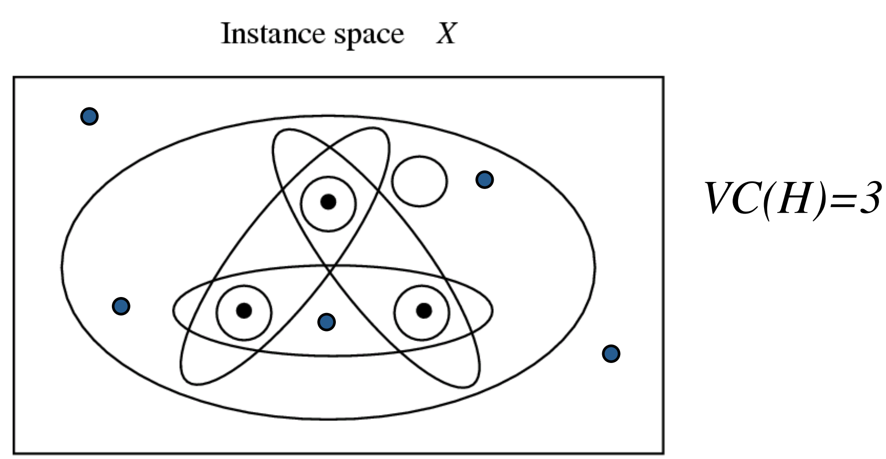
Each ellipse corresponds to a possible dichotomy

Positive: Inside the ellipse

Negative: Outside the ellipse

The Vapnik-Chervonenkis Dimension

Definition: The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.



Sample Complexity based on VC dimension

How many randomly drawn examples suffice to ε-exhaust $VS_{H,D}$ with probability at least (1-δ)?

ie., to guarantee that any hypothesis that perfectly fits the training data is probably (1-δ) approximately (ε) correct

$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

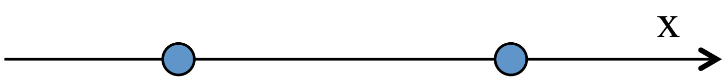
Compare to our earlier results based on $|H|$:

$$m \geq \frac{1}{\epsilon}(\ln(1/\delta) + \ln|H|)$$

VC dimension: examples

Consider 1-dim real valued input X, want to learn c:X→{0,1}

What is VC dimension of



- Open intervals:

  H1: if $x > a$ then $y = 1$ else $y = 0$

  H2: if $x > a$ then $y = 1$ else $y = 0$
  or, if $x > a$ then $y = 0$ else $y = 1$

- Closed intervals:

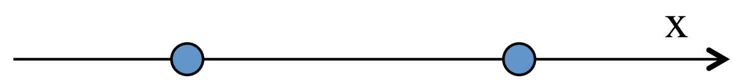
  H3: if $a < x < b$ then $y = 1$ else $y = 0$

  H4: if $a < x < b$ then $y = 1$ else $y = 0$
  or, if $a < x < b$ then $y = 0$ else $y = 1$

VC dimension: examples

Consider 1-dim real valued input X, want to learn c:X→{0,1}

What is VC dimension of



- Open intervals:

  H1: if $x > a$ then $y = 1$ else $y = 0$ VC(H1)=1

  H2: if $x > a$ then $y = 1$ else $y = 0$ VC(H2)=2
  or, if $x > a$ then $y = 0$ else $y = 1$

  H2 can perfectly handle if there is 2 X sample

- Closed intervals:

  H3: if $a < x < b$ then $y = 1$ else $y = 0$ VC(H3)=2

  H4: if $a < x < b$ then $y = 1$ else $y = 0$ VC(H4)=3
  or, if $a < x < b$ then $y = 0$ else $y = 1$

VC dimension: examples

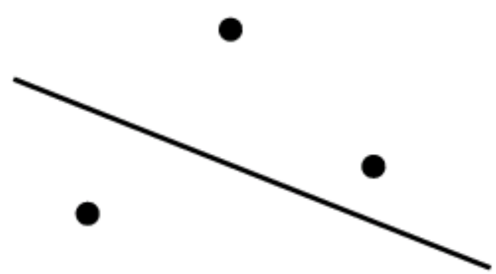
What is VC dimension of lines in a plane?

- $H_2 = \{ ((w_0 + w_1 x_1 + w_2 x_2) > 0 \rightarrow y=1) \}$

VC dimension: examples

What is VC dimension of

- $H_2 = \{ ((w_0 + w_1x_1 + w_2x_2)>0 \rightarrow y=1) \}$
  - $VC(H_2)=3$
- For H_n = linear separating hyperplanes in n dimensions, $VC(H_n)=n+1$



For any finite hypothesis space H, can you give an upper bound on VC(H) in terms of |H|? (hint: yes)

Assume VC(H) = K, which means H can shatter K examples.

For K examples, there are 2^K possible labelings. Thus, $|H| \geq 2^K$

Thus, $K \leq \log_2 |H|$

Tightness of Bounds on Sample Complexity

How many examples *m* suffice to assure that any hypothesis that fits the training data perfectly is probably (1-δ) approximately (ε) correct?

$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

How tight is this bound?

Tightness of Bounds on Sample Complexity

How many examples *m* suffice to assure that any hypothesis that fits the training data perfectly is probably (1-δ) approximately (ε) correct?

$$m \geq \frac{1}{\epsilon}(4 \log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

How tight is this bound?

Lower bound on sample complexity (Ehrenfeucht et al., 1989):

Consider any class C of concepts such that VC(C) > 1, any learner L, any $0 < \varepsilon < 1/8$, and any $0 < \delta < 0.01$. Then there exists a distribution and a target concept in C, such that if L observes fewer examples than

$$\max\left[\frac{1}{\epsilon}\log(1/\delta), \frac{VC(C)-1}{32\epsilon}\right]$$

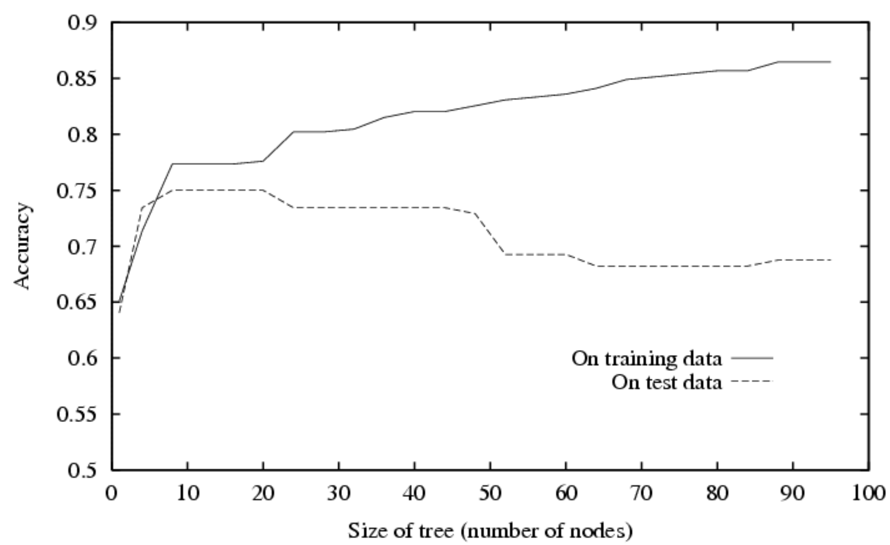
Then with probability at least δ, L outputs a hypothesis with $error_{\mathcal{D}}(h) > \epsilon$

Agnostic Learning: VC Bounds for Decision Tree

[Schölkopf and Smola, 2002]

With probability at least (1-δ) every $h \in H$ satisfies

$$error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}$$



What You Should Know

- Sample complexity varies with the learning setting
  - Learner actively queries trainer
  - Examples arrive at random

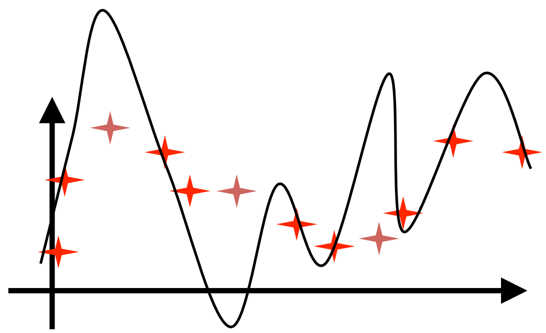
- Within the PAC learning setting, we can bound the probability that learner will output hypothesis with given error
  - For ANY consistent learner (case where $c \in H$)
  - For ANY "best fit" hypothesis (agnostic learning, where perhaps c not in H)

- VC dimension as a measure of complexity of H

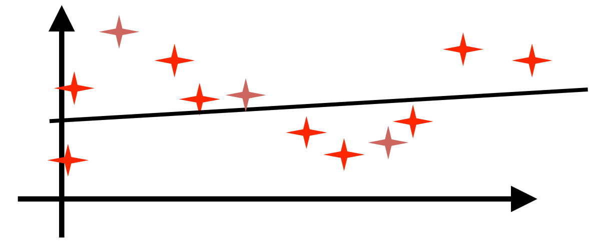
- Conference on Learning Theory: http://www.learningtheory.org
- Avrim Blum's course on Machine Learning Theory:
  - https://www.cs.cmu.edu/~avrim/ML14/

OVERFITTING, BIAS/VARIANCE TRADE-OFF

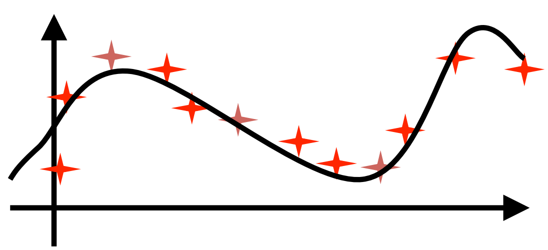
What is a good model?



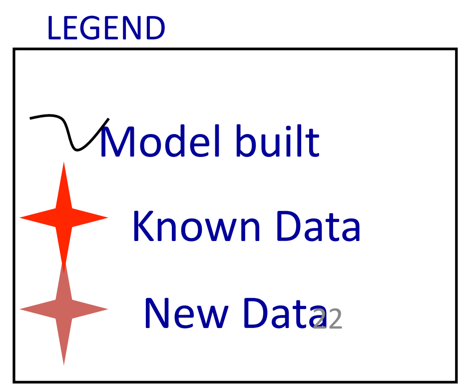
Low Robustness



Low quality /High Robustness

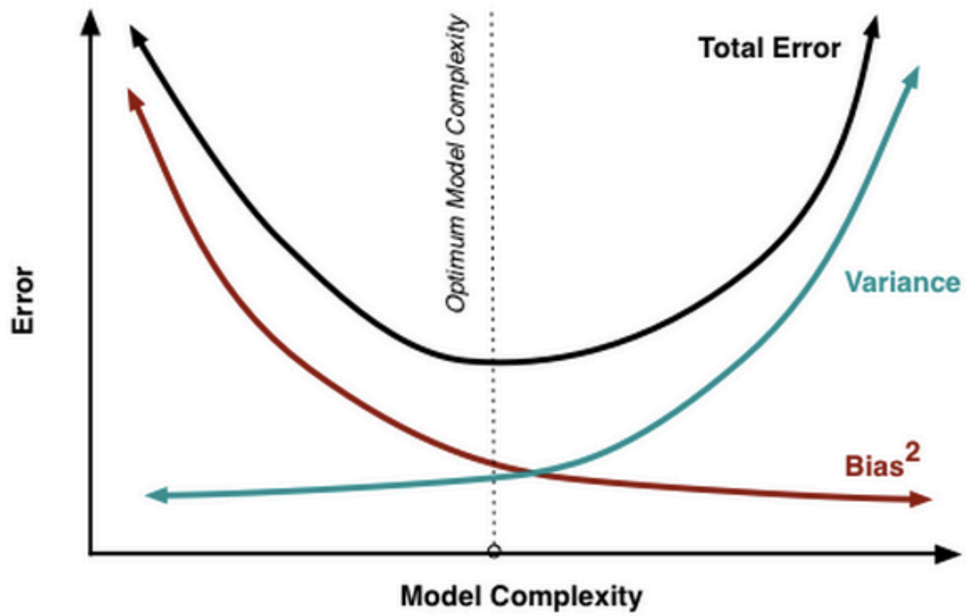


Robust Model



Two sources of errors

- Now let's look more closely into two sources of errors in an function approximator:



- In the following we show how bias and variance decompose

Expected loss, Bias/Variance Decomposition

- Let y be the **true** (target) output
- Let $h(x) = E[y|x]$ be the **optimal** predictor
- Let $f(x)$ our actual predictor, which will incur the following expected loss

$$E(f(x)-y)^2 = \int \left(f(x)-y\right)^2 p(x,y)dxdy$$

$$= \int \left(f(x)-h(x)+h(x)-y\right)^2 p(x,y)dxdy$$

$$= \int \left[\left(f(x)-h(x)\right)^2 + 2\left(f(x)-h(x)\right)\left(h(x)-y\right) + \left(h(x)-y\right)^2\right] p(x,y)dxdy$$

$$= \underbrace{\int \left(f(x)-h(x)\right)^2 p(x)dx}_{\text{The part we can influence by changing our predictor } f(x)} + \underbrace{\int \left(h(x)-y\right)^2 p(x,y)dxdy}_{\text{a noise term, and we can do no better than this. Thus it is a lower bound of the expected loss}}$$

Expected loss, Bias/Variance Decomposition

$$E(f(x)-y)^2 = \int \left(f(x)-h(x)\right)^2 p(x)dx + \int \left(h(x)-y\right)^2 p(x,y)dxdy$$

- $f(x;D)$: We will assume $f(x) = f(x|w)$ is a parametric model and the parameters w are fit to a training set D.
- $E_D[f(x;D)]$: The expected predictor over the multiple training datasets

Take the expectation over different datasets

$$E_D\left[\left(f(x;D)-h(x)\right)^2\right] = \underbrace{\left(E_D\left[f(x;D)\right]-h(x)\right)^2}_{\text{Bias}^2} + \underbrace{E_D\left[\left(f(x;D)-E_D\left[f(x;D)\right]\right)^2\right]}_{\text{Variance}}$$

Expected loss, Bias/Variance Decomposition

Proof:

$$E_D[\left(f(x;D) - h(x)\right)^2] = E_D[\left(f(x;D) - E_D\left[f(x;D)\right] + E_D\left[f(x;D)\right] - h(x)\right)^2]$$

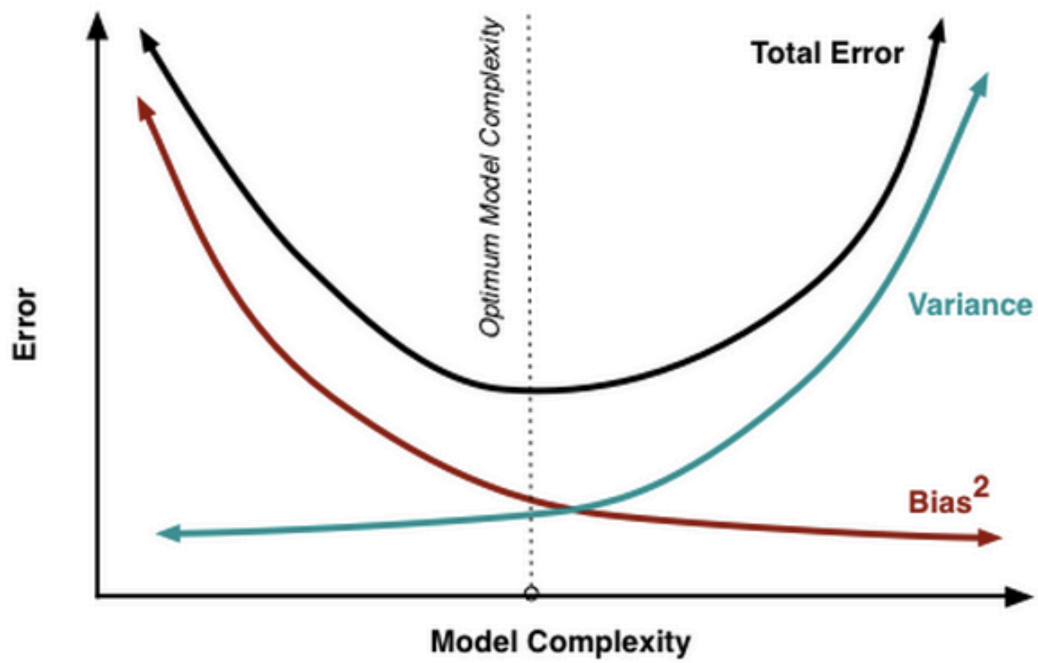
$$= E_D[\left(f(x;D) - E_D\left[f(x;D)\right]\right)^2 + \left(E_D\left[f(x;D)\right] - h(x)\right)^2$$

$$+ 2\left(f(x;D) - E_D\left[f(x;D)\right]\right)\left(E_D\left[f(x;D)\right] - h(x)\right)]$$

$$= \underbrace{\left(E_D\left[f(x;D)\right] - h(x)\right)^2}_{\text{Bias}^2} + \underbrace{E_D\left[\left(f(x;D) - E_D\left[f(x;D)\right]\right)^2\right]}_{\text{Variance}}$$

- Putting things together:

expected loss = $(\text{bias})^2$ + variance + noise



expected loss = $(\text{bias})^2$ + variance + noise

Regularized Regression

- Recall linear regression: $$\mathbf{y} = \mathbf{X}^T\beta + \varepsilon$$

$$\beta^* = \arg\max_\beta (\mathbf{y} - \mathbf{X}^T\beta)^T(\mathbf{y} - \mathbf{X}^T\beta)$$

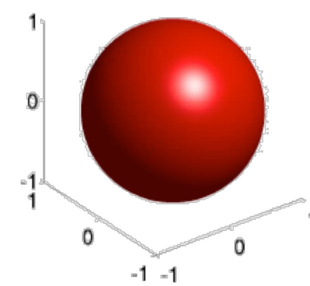
$$= \arg\max_\beta \|\mathbf{y} - \mathbf{X}^T\beta\|^2$$

- Regularized LR:
  - L2-regularized LR:

$$\beta^* = \arg\max_\beta \|\mathbf{y} - \mathbf{X}^T\beta\|^2 + \lambda\|\beta\|$$

where

$$\|\beta\| = \sum_i \beta_i^2$$

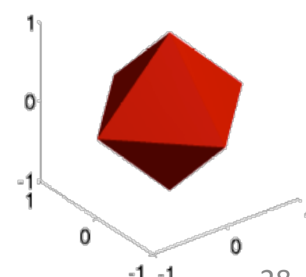


  - L1-regularized LR:

$$\beta^* = \arg\max_\beta \|\mathbf{y} - \mathbf{X}^T\beta\|^2 + \lambda|\beta|$$

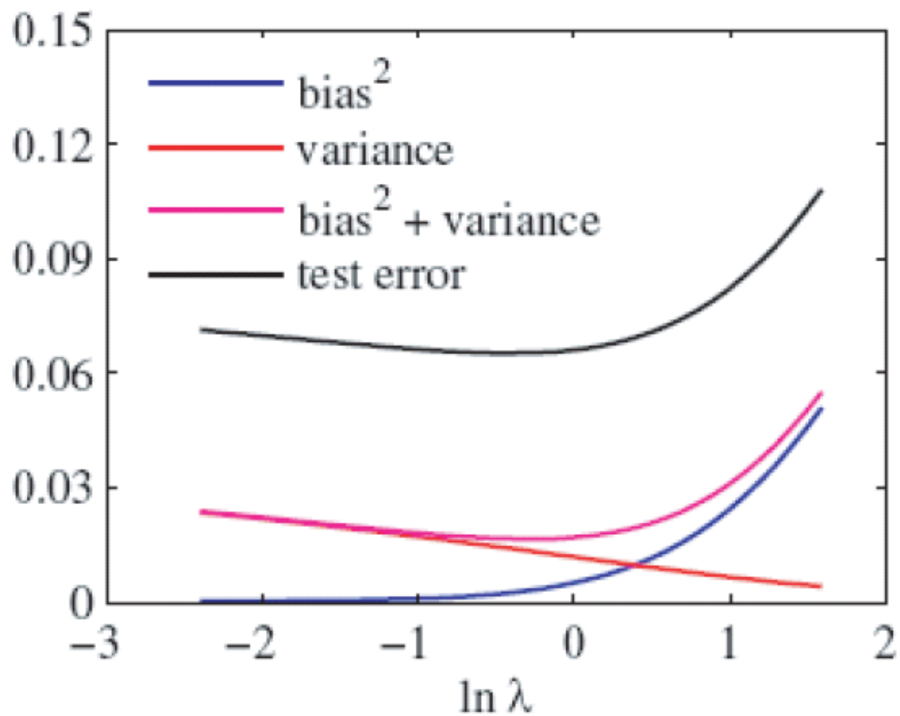
where

$$|\beta| = \sum_i |\beta_i|$$



λ controls bias/variance trade off

Bias^2+variance vs regularizer



- Bias^2+variance predicts (shape of) test error quite well.
- However, bias and variance cannot be computed since it relies on knowing the true distribution of x and y (and hence $h(x) = E[y|x]$).

Bayes Error Rate

- Fundamental performance limit for classification problem
- A lower bound on classification performance of *any* algorithms on a given problem
    - i.e., Error rate of the optimal decision rule

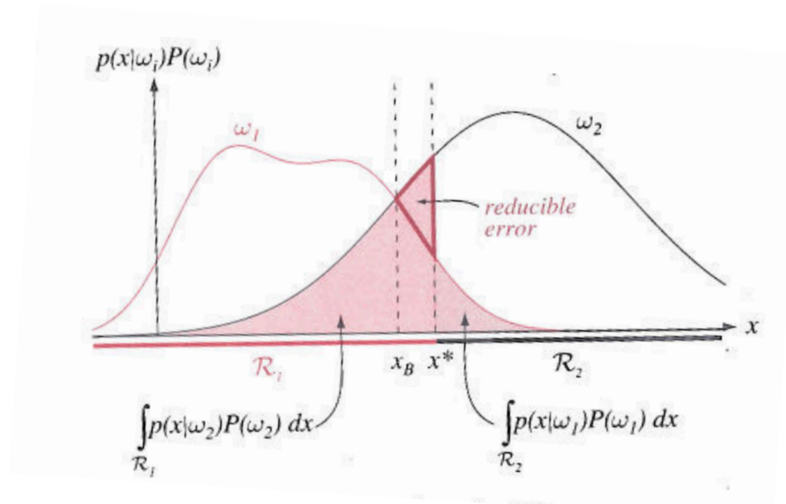
Bayes Error Rate: Two Class

- For a two-class classification problem
  - x is input feature vector, and ω_1, ω_2 are two classes
  - Then, Bayes optimal decision rule is
    - Classify as ω_1 if

$$P(\mathbf{x} \mid \omega_1)P(\omega_1) > P(\mathbf{x} \mid \omega_2)P(\omega_2)$$

    - Classify as ω_2 if

$$P(\mathbf{x} \mid \omega_1)P(\omega_1) < P(\mathbf{x} \mid \omega_2)P(\omega_2)$$



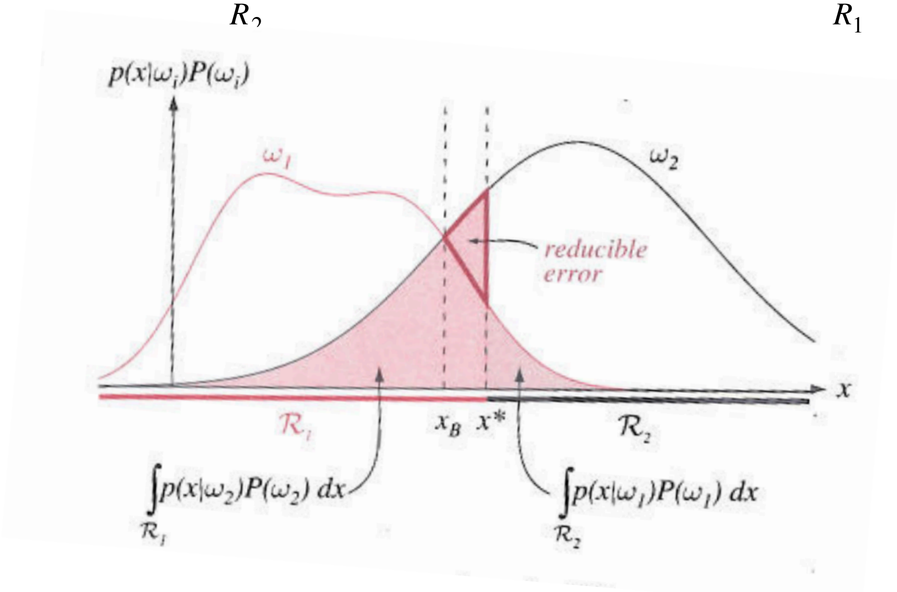
Bayes Error Rate: Two Class

- For a two-class classification problem
  - x is input feature vector, and ω_1, ω_2 are two classes
  - Given this optimal decision rule, the error rate is

$$P(error) = P(\mathbf{x} \in R_2, \omega_1) + P(\mathbf{x} \in R_1, \omega_2)$$

$$= P(\mathbf{x} \in R_2 \mid \omega_1)P(\omega_1) + P(\mathbf{x} \in R_1 \mid \omega_2)P(\omega_2)$$

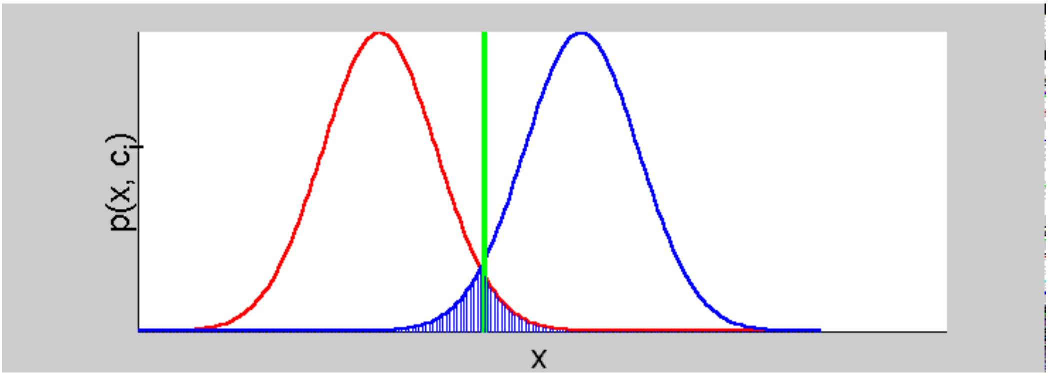
$$= \int_{R_2} P(\mathbf{x} \in R_2 \mid \omega_1)P(\omega_1)d\mathbf{x} + \int_{R_1} P(\mathbf{x} \in R_1 \mid \omega_2)P(\omega_2)d\mathbf{x}$$



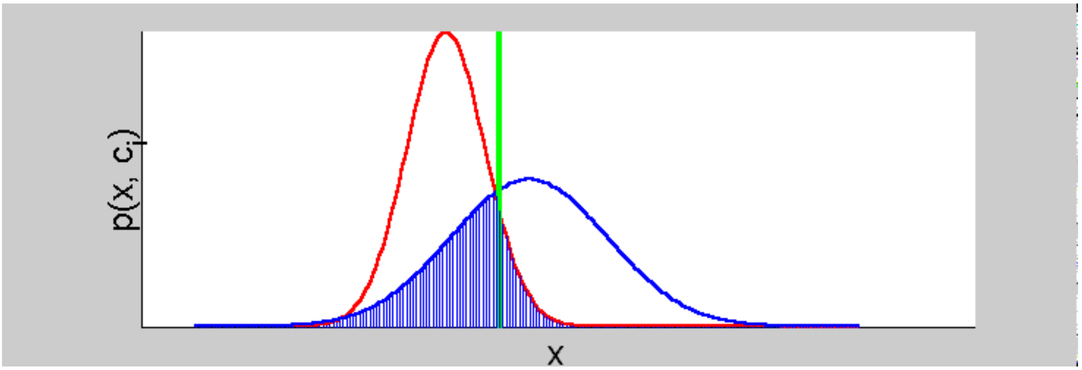
Bayes error rate gives the irreducible error: fundamental property of the problem, not the classifier

Classification Example

- Simple problem



- Hard problem



Bayes Error Rate: Multiple Classes

- For c-class classification

$$P(correct) = \sum_{i=1}^{c} P(\mathbf{x} \in R_i, \omega_i)$$

$$= \sum_{i=1}^{c} P(\mathbf{x} \in R_i \mid \omega_i) P(\omega_i)$$

$$= \sum_{i=1}^{c} \int_{R_2} P(\mathbf{x} \in R_i \mid \omega_i) P(\omega_i) d\mathbf{x}$$

$$P(error) = 1 - P(correct)$$

Summary

- Overfitting
- Bias-variance decomposition
- Bayes Error Rate