
Model Selection:

choosing the best classifier for your task

EE511 Statistical Learning

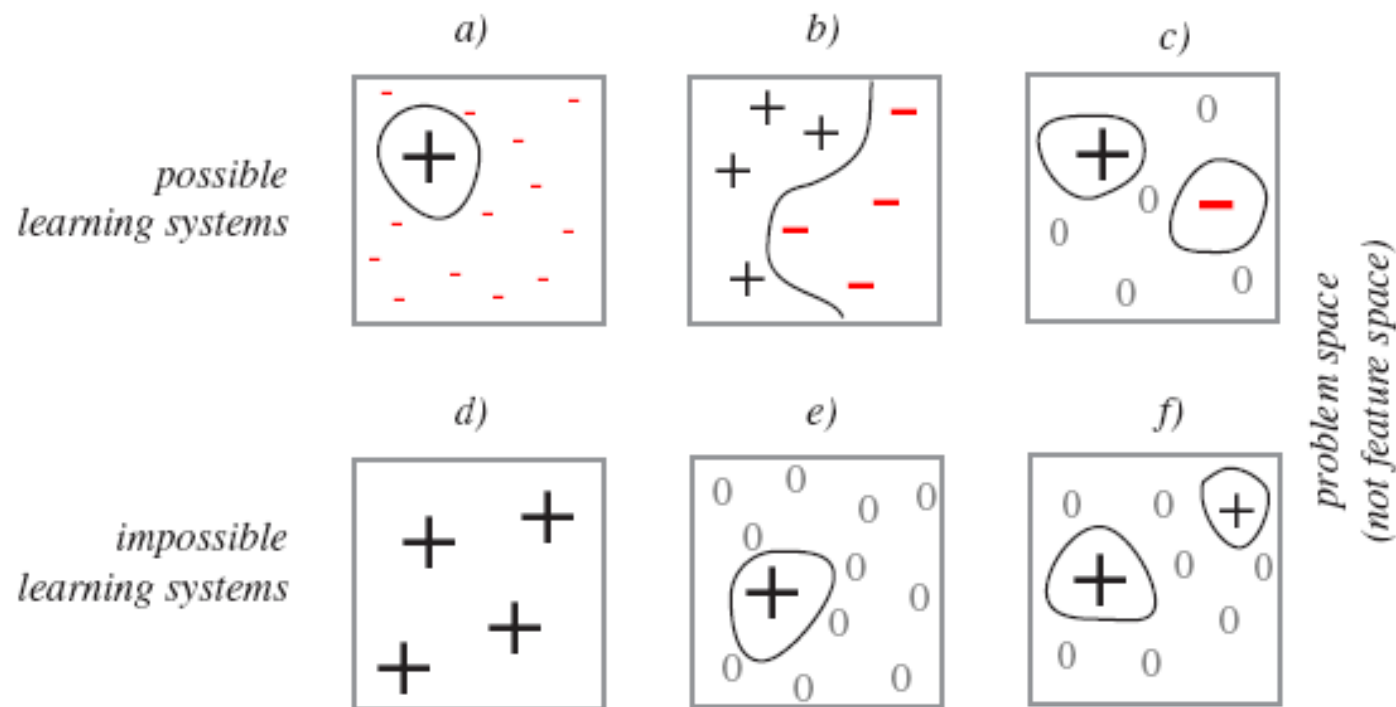
May 6, 2008

Kevin Duh

No Free Lunch Theorem

- Is there a learning algorithm that is inherently better than others?
 - Better = Lower generalization error
- “No Free Lunch Theorem” says: **NO!**
 - No algorithm is universally good for all problems.
 - Averaged over all problems, all algorithms have equal error
 - This is true independent of $P(x)$ and data size

No Free Lunch: “Law of conservation”



- If a learning algorithm performs above-average in some problems, it will perform below-average in others (Figure from Duda, Hart, Stork, “Pattern Classification”)

Model Selection

- For a **particular task/dataset**, an optimal algorithm may exist
- The goal of model selection:
 - Using your training data, find the model that is expected to generalize best
 - Model = different learning algorithms, different parameter settings of the same algorithm
 - Here, we'll use “classifier” and “model” interchangeably
- Related concepts:
 - Model assessment: Given a model, estimate its generalization error, error bars, etc.
 - Model averaging: Combine models as opposed to selecting the best model

Today's Agenda

1. Re-visit Bias-Variance Tradeoff
2. General concepts in Model Selection
3. Extra-sample methods
 - Cross-Validation
 - Leave-one-out error
4. Intra-sample methods
 - Bayesian model selection & BIC
 - AIC
 - MDL
5. Vapnik-Chervonekis Theory

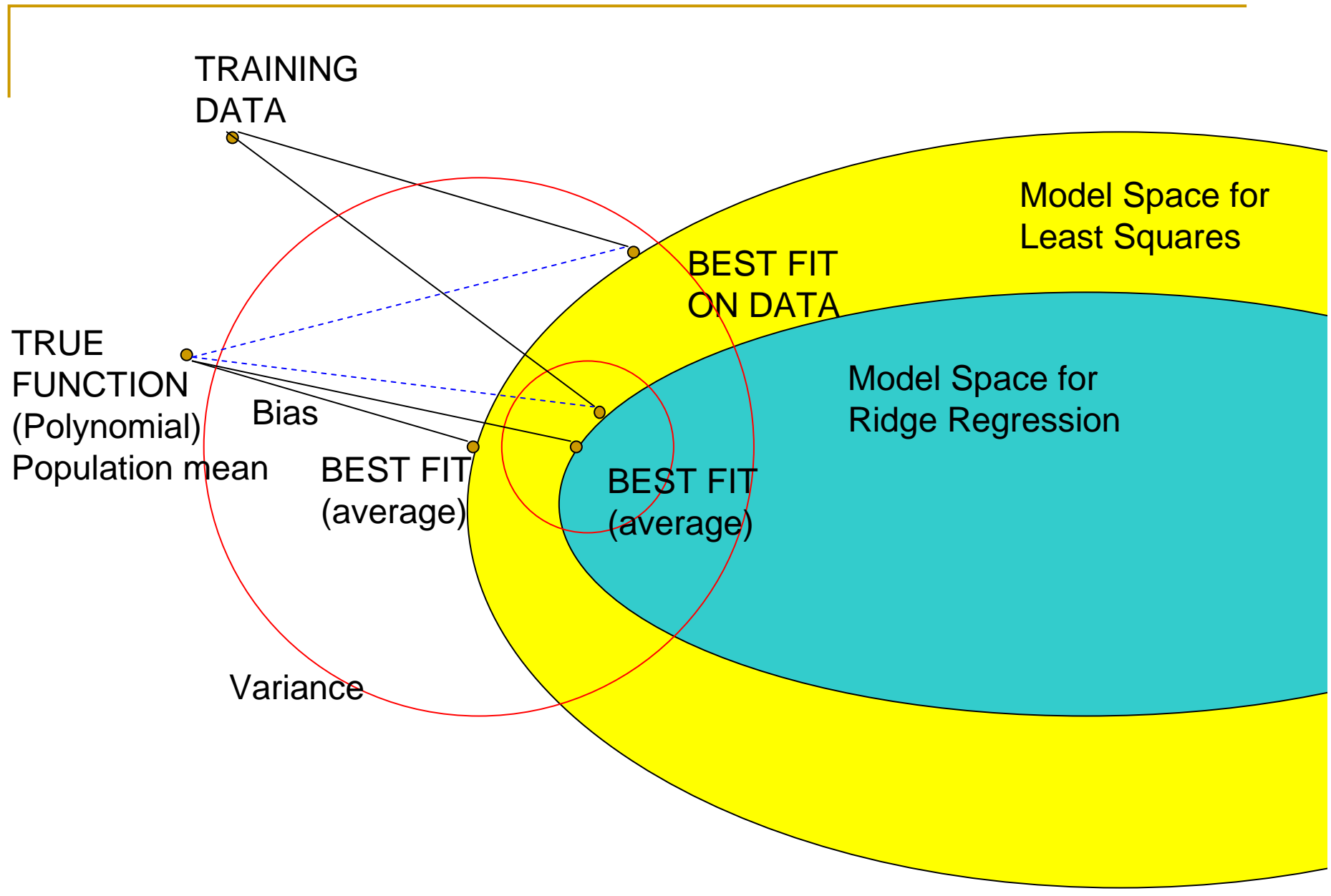
Recurring theme:
Complexity of classifier

Bias-Variance Tradeoff for Regression

- Mean Squared Error = bias² + variance

$$E[(f(x) - y)^2] = E[f(x) - y]^2 + E[(f(x) - E[f(x)])^2]$$

- Low bias: on average, $f(x)$ is close to truth
 - Low variance: $f(x)$ does not change much as training data varies
-
- Model selection: find the best balance
 - To reduce bias, increase model complexity (generally)
 - To reduce variance, decrease model complexity (generally)



Bias-Variance Tradeoff for Classification

- For 0-1 loss with $y = \{0, 1\}$, $\Pr(y=1|x) = f(x)$:

$$\Pr(f(x) \neq y) = \Phi[\underbrace{\text{sign}(1/2 - y)E[f(x) - 1/2]}_{\text{Boundary bias}} \underbrace{\text{Var}[f(x)]^{-1/2}}_{\text{Variance}}]$$

$$\Phi[t] = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-u^2/2} du$$

- Different from regression case. Here:
 - Nonlinear & multiplicative interaction between bias/variance
 - If bias is negative, low variance reduces $\Pr(\text{error})$
 - If bias is positive, high variance reduces $\Pr(\text{error})$
 - **Variance dominates bias**
 - **One reason why classifiers care so much about complexity**

Today's Agenda

1. Re-visit Bias-Variance Tradeoff
2. **General concepts in Model Selection**
3. Extra-sample methods
 - Cross-Validation
 - Leave-one-out error
4. Intra-sample methods
 - Bayesian model selection & BIC
 - AIC
 - MDL
5. Vapnik-Chervonekis Theory

General concepts in Model Selection

- Basic ingredients:

- Goodness-of-fit

- Model complexity

} Methods vary by different definitions of these

- Training a classifier = optimizing “goodness-of-fit”

- Objective criteria is defined

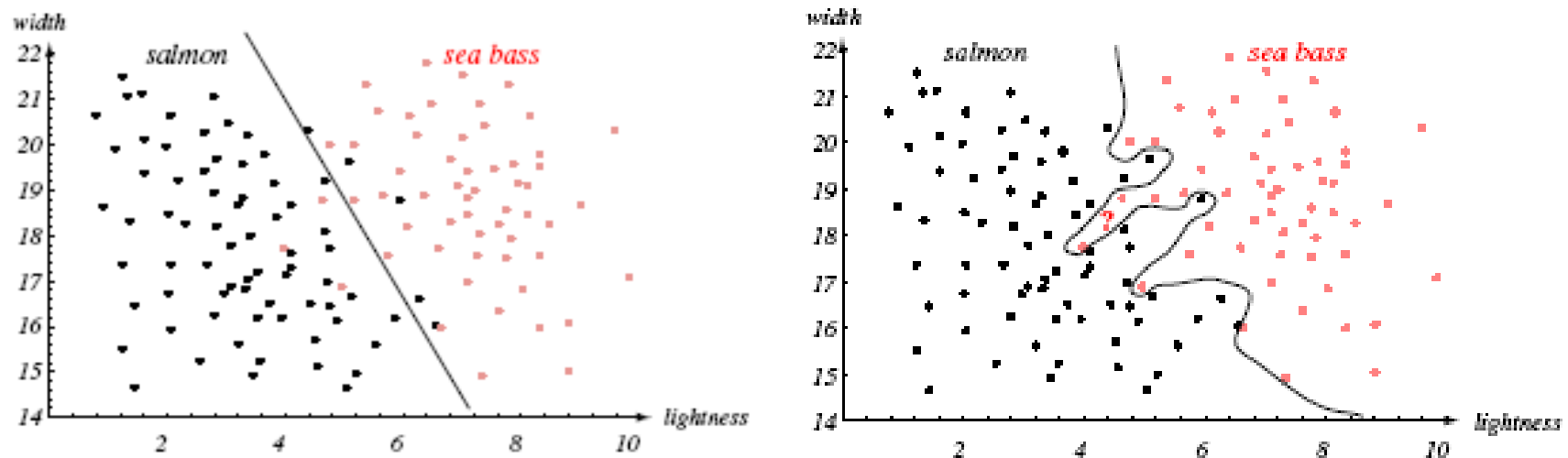
- Learning algorithm implements an optimization procedure to pick classifier that best fits the data

- This “selects” the best classifier among all classifiers of the same class, where “best” = attains best objective value

- **Issue: Classifiers from different classes cannot be compared using the objective**

Why can't we compare across different classes of classifiers?

- Task: binary classification
- Training objective: min error on training set
- Two different classes of classifiers: linear vs. very wiggly



- Which achieves better objective? Which will you use?

Occam's Razor

- William of Ockham (14th century logician):
 - *“entia non sunt multiplicanda praeter necessitatem”*
 - “entities should not be multiplied beyond necessity”
- In the case of statistical learning:
 - Prefer models that fit the data but have minimal complexity

Optimism of the Training Error Rate

- Goal: Low test (generalization) error

$$TestErr = E \left[L \left(Y, \hat{f} (X) \right) \right]$$

- Typically: training error rate < test error
 - Same data is used to fit the model & assess its error

$$TrainErr = \frac{1}{N} \sum_{i=1}^n L \left(y_i, \hat{f} (x_i) \right)$$

- Test error is a kind of extra-sample error
 - Features in the test set are not observed in the training set
 - Cross-validation methods estimate this test error using a held-out set

In-sample error

1. Assuming a data generation process: $y = f(x) + \text{noise}$:

$$Err_{in} = \frac{1}{N} \sum_{i=1}^N E_y E_{Y^{new}} L(Y_i^{new}, \hat{f}(x_i))$$

Y^{new} = **new** response values at each of training points x_i , $i=1, 2, \dots, N$

2. Define optimism $op \equiv Err_{in} - E_y(\overline{err})$

For squared error, 0-1 loss functions, it can be shown:
i.e. if we try too hard to fit y , then optimism is high $op = \frac{2}{N} \sum_{i=1}^N Cov(\hat{y}_i, y_i)$

3. Model selection formula:

AIC, BIC, MDL, etc. can be seen as variations of this.

$$Err_{in} = E_y(\overline{err}) + \frac{2}{N} \sum_{i=1}^N Cov(\hat{y}_i, y_i)$$

What are properties of a good model selection method?

- Accurately estimates generalization error
 - Has low bias and variance itself
 - Consistency:
 - Selects the true model (assuming it's in the class of classifiers considered) as number of training samples increases to infinity
 - Computable
 - No tunable parameters
 - Widely applicable
-

Today's Agenda

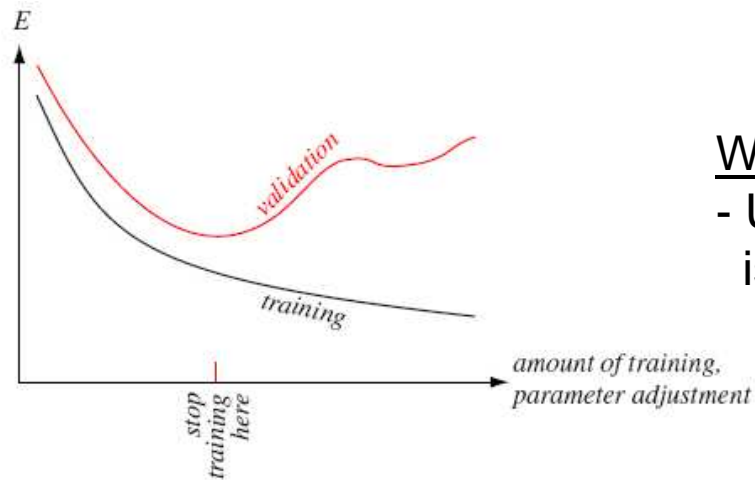
1. Re-visit Bias-Variance Tradeoff
2. General concepts in Model Selection
3. **Extra-sample methods**
 - **Cross-Validation**
 - **Leave-one-out error**
4. Intra-sample methods
 - Bayesian model selection & BIC
 - AIC
 - MDL
5. Vapnik-Chervonekis Theory

Extra-sample methods

- Directly estimate test error by using a held-out (validation) set
 - Validation, cross-validation, Leave-one-out error
 - Bootstrap
- Assumptions:
 - Validation set is similar to test set
 - Validation error is sufficiently accurate
- These methods are often used in practice
 - Usually achieves good estimates
 - Does not assume anything about the model (parametric, non-parametric) or the task (classification, regression, density estimation)

Hold-out set for model selection

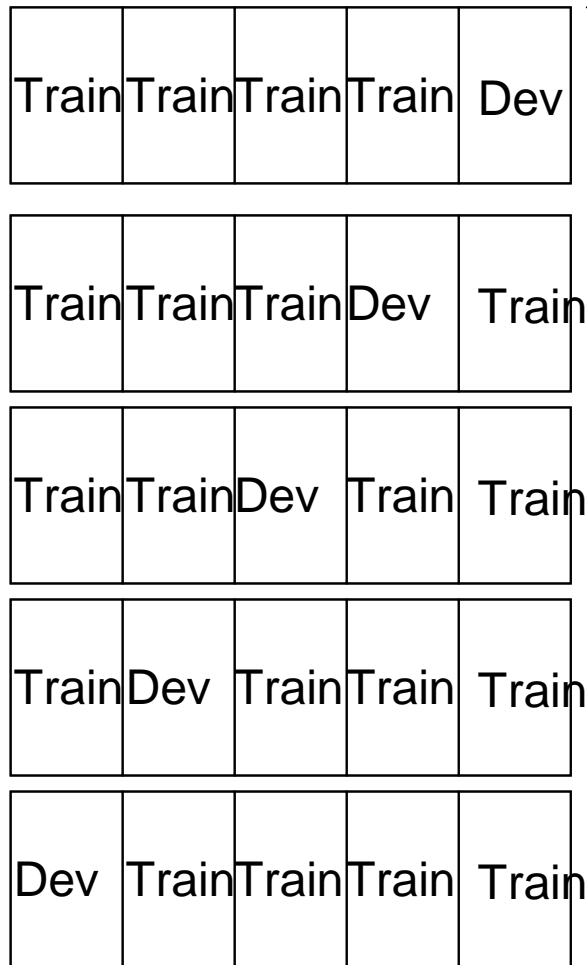
- Use a development (validation) set to choose parameters/models



What's the proportion of dev set?

- Usually small (10-20%), since training data is needed to estimate many model parameters

K-Fold Cross-Validation



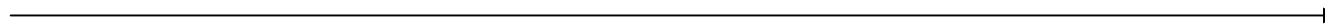
To calculate a model's performance, average the performance of each fold

- More robust than single dev set.
- Usually 5-fold or 10-fold used;
- Leave-one-out error in the extreme case
- Stratified CV: maintain label proportions

How many folds in K-fold cross-validation?

Less Variance

Less Bias



LARGER K

If $K=N$ (leave-one-out error), we get approximately unbiased estimate for test error

- but high variance due to N very similar training sets
- high computational burden (but some algorithms have clever tricks to do this)

For low K , bias may be a problem depending on size of training set

Learning Curve

Large bias here
(overly pessimistic about true error)

If we operate at this range or above,
Cross-validation error has small bias

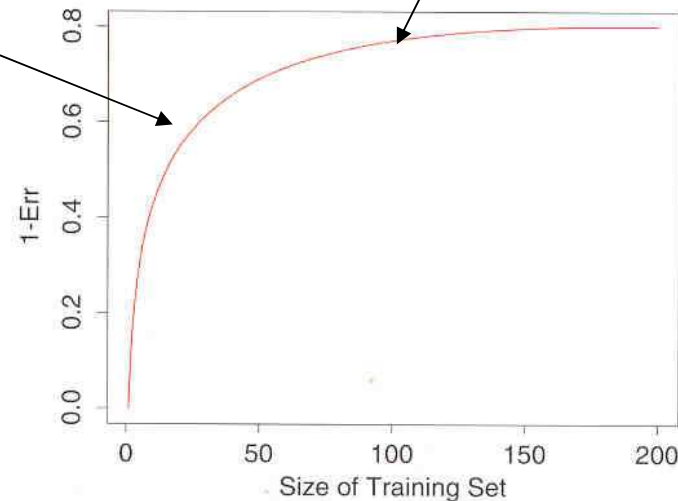


FIGURE 7.8. Hypothetical learning curve for a classifier on a given task; a plot of $1 - \text{Err}$ versus the size of the training set N . With a dataset of 200 observations, fivefold cross-validation would use training sets of size 160, which would behave much like the full set. However, with a dataset of 50 observations fivefold cross-validation would use training sets of size 40, and this would result in a considerable overestimate of prediction error.

- Figure from *The Elements of Statistical Learning* by Hastie, Tibshirani and Friedman.

Covariance matrix of cross-validation errors has a block structure

Corollary 1 *The covariance matrix Σ of cross-validation errors $\mathbf{e} = (e_1, \dots, e_n)'$ has the simple block structure depicted in Figure 2: 1) all diagonal elements are identical $\forall i, \text{Cov}(e_i, e_i) = \text{Var}[e_i] = \sigma^2$; 2) all the off-diagonal entries of the K $m \times m$ diagonal blocks are identical $\forall (i, j) \in T_k^2 : j \neq i, T(j) = T(i), \text{Cov}(e_i, e_j) = \omega$; 3) all the remaining entries are identical $\forall i \in T_k, \forall j \in T_\ell : \ell \neq k, \text{Cov}(e_i, e_j) = \gamma$.*

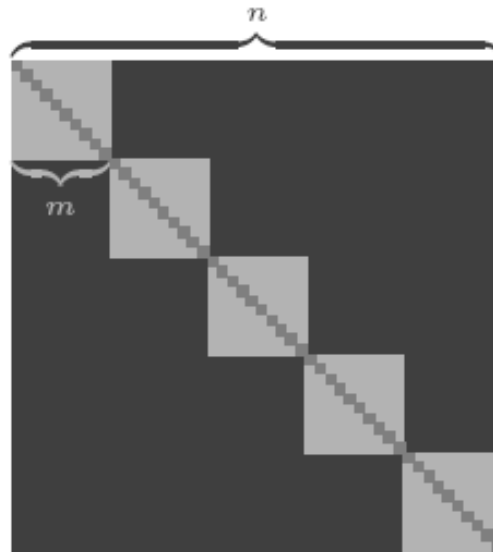


Figure 2: Structure of the covariance matrix.

Summary of Extra-sample methods

- K-fold Cross-validation is effective in practice
 - K=5 or 10 is usually good, but do think about bias-variance questions
 - It's not a cure-all, or else it'd violate "No Free Lunch"
- Bootstrap is an alternative
 - (not discussed here)
 - Basic idea: resamples the training set with replacement to reduce variance

Today's Agenda

1. Re-visit Bias-Variance Tradeoff
2. General concepts in Model Selection
3. Extra-sample methods
 - Cross-Validation
 - Leave-one-out error
4. Intra-sample methods
 - Bayesian model selection & BIC
 - AIC
 - MDL
5. Vapnik-Chervonekis Theory

Bayesian Model Selection & BIC (1 / 5)

- We'll derive the BIC using Bayesian principles
- Notation:
 - D = data, w = parameters, H_i = model i
- Two levels of being Bayesian:
 - Bayesian parameter estimation
$$P(w | D, H_i) = \frac{P(D | w, H_i) \cdot P(w | H_i)}{P(D | H_i)}$$
 - Posterior = Likelihood x Prior / Evidence
 - Bayesian model selection

Usually ignored;
doesn't affect solution

Bayesian model selection (2/5)

- Bayesian parameter estimation

$$P(w | D, H_i) = \frac{P(D | w, H_i) \cdot P(w | H_i)}{P(D | H_i)}$$

- Bayesian model selection

$$P(H_i | D) = \frac{P(D | H_i) \cdot P(H_i)}{P(D)}$$

- Evidence $P(D|H_i)$ is important
- Select models H_1 vs H_2 based on posterior odds:

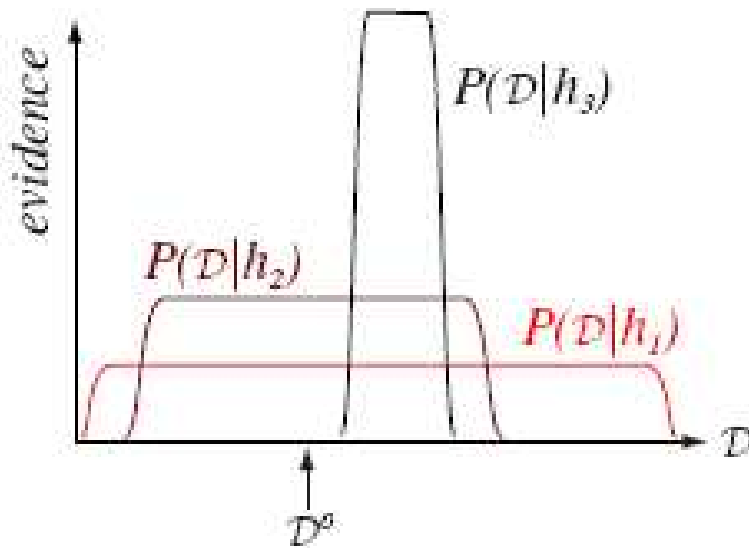
$$\frac{P(H_1 | D)}{P(H_2 | D)} = \frac{P(H_1)}{P(H_2)} \cdot \frac{P(D | H_1)}{P(D | H_2)}$$

Bayes Factor

Bayesian model selection (3/5)

You might imagine we need to encode the priors to penalize complex models...
But it turns out it is captured in evidence $P(D|H_i)$ automatically

$$\frac{P(H_1 | D)}{P(H_2 | D)} = \frac{P(H_1)}{P(H_2)} \cdot \frac{P(D | H_1)}{P(D | H_2)}$$



Complex H will have large support
because it can fit many datasets
- That implies it is less peaked

Bayesian model selection (4/5): How to evaluate the evidence?

- To compute evidence, integrate over all parameters:

$$P(D | H_i) = \int P(D | w, H_i) \cdot P(w | H_i) dw$$

- Laplacian Approximation: suppose $P(w|D, H_i)$ is peaky at MAP solution

$$P(D | H_i) \approx P(D | w_{MAP}, H_i) \cdot P(w_{MAP} | H_i) \sigma_{w|D}$$

← Width of peak

$$= P(D | w_{MAP}, H_i) \cdot \left(\frac{1}{\sigma_w} \right)$$

← Uniform prior over all w

Best fit likelihood

Occam Factor

- Occam Factor = factor in which H_i 's hypothesis space collapses after seeing data.
 - Penalizes complex models (i.e. high σ_w)
 - Penalizes models that fit too much to the data (i.e. low $\sigma_{w|D}$)

BIC: Bayesian Information Criteria (5/5)

- BIC can be derived from a Gaussian approximation of the evidence term
 - See Raftery (1995) “Bayesian model selection in social research”, Social Methodology, Vol 25

- Basic Form:

$$BIC = -2 \cdot (\log lik) + (\overset{\text{\#samples}}{\log N}) \cdot \overset{\text{\#parameters}}{d}.$$

- Choose the model that has the lowest BIC
- BIC tends to penalize complex models heavily
- Note dependence on N

Akaike Information Criteria (AIC)

- AIC is another estimate of in-sample prediction error
 - Derived from information theoretic arguments

$$\text{For } N \rightarrow \infty : \quad -2E \left[\log \Pr_{\hat{\theta}} (Y) \right] \approx -\frac{2}{N} E [\log \text{lik}] + 2 \frac{d}{N}$$

$\Pr_{\theta} (Y)$... family density for Y (containing the true density)

$\hat{\theta}$... ML estimate of θ

$$\log \text{lik} = \sum_{i=1}^N \log \Pr_{\hat{\theta}} (y_i)$$

Maximized log-likelihood due to ML estimate of theta

AIC or BIC?

- Recall we choose model with small BIC/AIC

$$BIC = -2 \cdot (\log lik) + (\log N) \cdot d$$

$$AIC = -\frac{2}{N} (\log lik) + 2\left(\frac{d}{N}\right)$$

- BIC is asymptotically consistent. AIC is not.
- AIC tends to choose more complex models as N increases
- BIC usually favors simpler models, due to strong penalty
- Asymptotically, AIC = leave-one-out; Asymptotically, BIC = cross validation with a particular K. To see this, note:

$$\log P(D | H_i) = \log P(d_1 | H_i) + \log P(d_2 | d_1 | H_i) + \log P(d_N | d_1, d_2, \dots, d_{N-1}, H_i)$$

Cross validation examines the average of this term over reorderings of data

MDL: Minimum Description Length

- MDL can be derived from Bayesian model selection and vice versa. (But originally derived independently)
 - Replaces probability of events by code length required to communicate the event
 - $L(x) = -\log_2(P(x))$. \leq length of x in bits
- MDL criteria chooses the model with the least bits.

$$MDL = -\log P(H) - \log(P(D | H) \delta D)$$

$= -\log P(H | D) + \text{constant}$

The diagram shows two arrows pointing from labels below to terms in the equation above. One arrow points from 'Model block' to $-\log P(H)$. Another arrow points from 'Data block' to $-\log(P(D | H) \delta D)$.

- Other issues: precision of parameters

Summary of In-Sample Estimates

- AIC, BIC, MDL:
 - Many possible derivations
 - All contain 2 terms: goodness-of-fit & complexity
- AIC/BIC/MDL vs. Cross-Validation?
 - Personally I'd choose CV (more practical, applies to more kinds of models) unless there's good theoretical reasons to motivate AIC/BIC/MDL, etc.

Today's Agenda

1. Re-visit Bias-Variance Tradeoff
2. General concepts in Model Selection
3. Extra-sample methods
 - Cross-Validation
 - Leave-one-out error
4. Intra-sample methods
 - Bayesian model selection & BIC
 - AIC
 - MDL
5. Vapnik-Chervonekis Theory

Vapnik-Chervonenkis Theory

- VC Theory: attempts to explain learning from the statistical point of view
- Several important concepts/results
 - ❑ VC dimension & generalization error bounds
 - ❑ Structural risk minimization
 - ❑ Support vector machines

We'll mainly discuss this here in the context of model selection but note that it is a more general theory of learning (not just model selection)

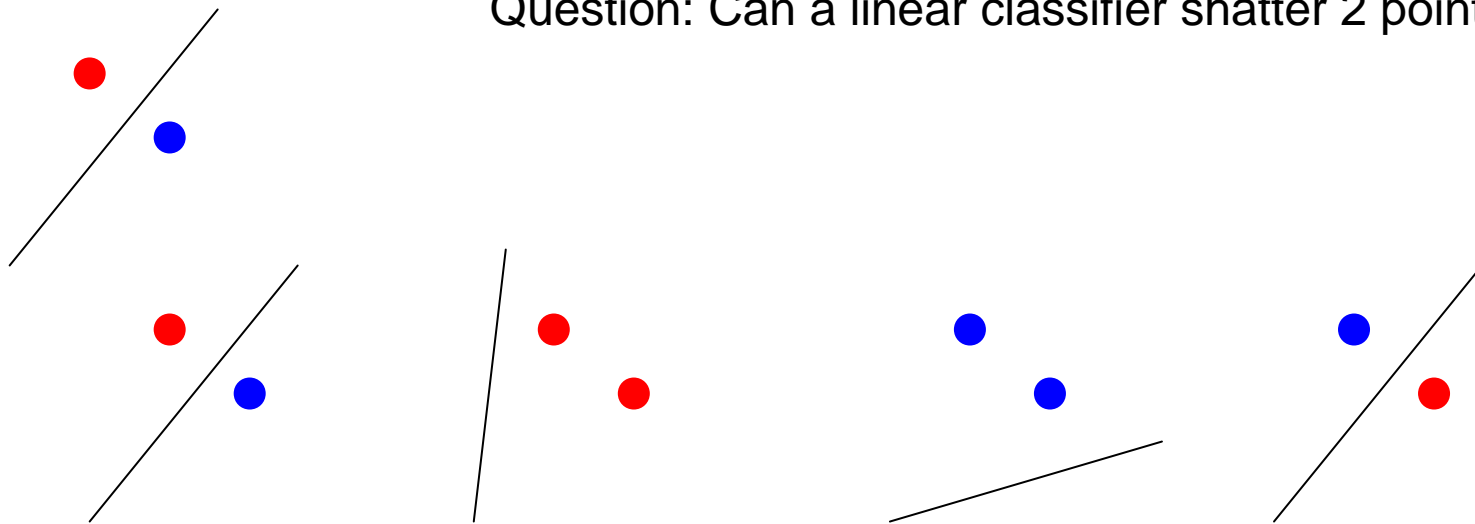
VC Dimension

- Recall that BIC/AIC/MDL counts the number of parameters in the penalty term
 - For some models, #parameters is clear, but not for others (esp. non-linear models)
- VC dimension measures complexity of classifier class in relation to a general training set of m samples
 - Class of (possibly infinite) functions: $\{f(x,z)\}$ is indexed by z
 - The complexity of this class is related to how many points (m) that can be **shattered (i.e. 2^m labels of m points being perfectly separable)**

VC Dimension Example: Shattering

- $\{f(x,z)\}$ can shatter m points IFF for every possible training set of the form $\{(x_1,y_1),(x_2,y_2),\dots,(x_m,y_m)\}$, there exists some z such that $f(x,z)$ gets zero training error

Question: Can a linear classifier shatter 2 points in 2D?



Pop Quiz

- Can a linear classifier shatter 3 points in 2D?
- Can a linear classifier shatter 4 points in 2D?
- VC Dimension = maximum number of points that can be shattered by a class of classifiers
- What's the VC Dimension of a linear classifier of k -dimensions?
- What function has infinite VC Dimension?

VC Dimension and Shattering

(another view)

- Suppose we have a class of functions $\{f(x,z)\}$ that has **infinite** number of elements
- But: Given **m** points, there are only **2^m** possible sub-classes of distinct functions
- Shattering coefficient measures the number of ways a function class can separate samples
 - i.e. it's the number of different outputs (y_1, y_2, \dots, y_m) that can be achieved by the function class
 - If shattering coefficient = 2^m , all possible separations can be implemented, and the function class is said to be able to shatter m points
- NOTE: this means there exist a set of m patterns that can be separated in all possible ways (it does not mean all sets of m patterns need to be shattered)

VC Generalization Bound

- Vapnik showed that with probability $1-\eta$

$$Err_{true} \leq Err_{train} + \frac{\varepsilon}{2} \left(1 + \sqrt{1 + \frac{4 \cdot Err_{train}}{\varepsilon}} \right)$$

$$\text{where } \varepsilon = a_1 \frac{h[\log(a_2 N/h) + 1] - \log(\eta/4)}{N}$$

$h = VC$ dimension

These kind of bounds can be used for model selection, as well.

But like AIC/BIC/MDL, your mileage varies:

- some have stricter bound, some easier computable

VC Theory (Big picture)

- Main question: when does learning work?
 - Given a class of functions, we minimize training error on m points find the best f .
 - How do we know that this f leads to the lowest attainable test error as m goes to infinity (consistency)?
 - ie. f is as good as directly minimizing test error, in the limit
 - Without restricting the set of functions, it turns out that empirical risk minimization is **inconsistent!**
 - Can you think of an example?
-

How to derive a VC bound

- Theorem (Vapnik & Chervonenkis): one side uniform convergence in probability,

$$\lim_{x \rightarrow \infty} P\{\sup_{f \in \mathcal{F}} (Err_{true}[f] - Err_{train}[f]) > \epsilon\} = 0, \forall \epsilon > 0$$

Is a necessary and sufficient condition for consistency of empirical risk minimization

- We can bound LHS as a function of VC dimension and sample size (many tricks here)

$$P\{\sup_{f \in \mathcal{F}} (Err_{true}[f] - Err_{train}[f]) > \epsilon\} \leq \text{function}(VC \text{ dim}, m, \epsilon)$$

- Using confidence intervals, re-express the above as:

$$Err_{true} \leq Err_{train} + \text{function}(\delta, VC \text{ dim}, m, \epsilon) \text{ w/ prob } \delta$$

Overall Summary

- No Free Lunch
- Central issues:
 - How to compare models of different complexity
 - Occam's Razer
- BIC, AIC, MDL:
 - different goodness-of-fit & complexity terms
 - Bayesian model selection as motivation for BIC
- Direct estimate of test error by hold-out
 - Cross validation, Leave-one-out
- Bias-variance for classifiers, Bias-variance for estimators of test error
- VC dimension as an measure of effective classifier complexity
- VC theory: learning is all about restricting the complexity of class of classifiers (without any restriction, learning is impossible)