Foundation of Machine Learning
(CSE4032)
Lecture 09: Model Assessment and Selection

Dr. Kundan Kumar
Associate Professor
Department of ECE

Faculty of Engineering (ITER)
S'O'A Deemed to be University, Bhubaneswar, India-751030

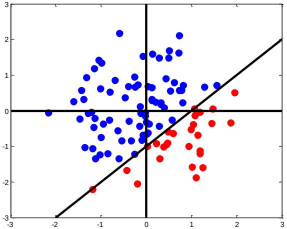
Topics to be covered

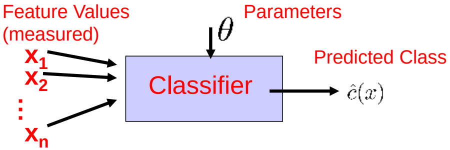
- Bias, Variance and Model Complexity (covered in previous lectures)
- Model Selection
  - Estimating the performance of different models in order to choose the best one. Eg. AIC, BIC
- Model assessment (covered in previous lectures)
  - having chosen a final model, estimating its prediction error (generalization error) on new data. e.g. Confusion matrix, Accuracy, TPR, FPR, etc
- Training Error Rate (covered in previous lectures)
- Prediction Error (covered in previous lectures)
- Vapnik–Chervonenkis Dimension (way of measuring the complexity)

Vapnik–Chervonenkis Dimension

Learners and Complexity

- We've seen many versions of underfit/overfit trade-off
  - Complexity of the learner
  - Representational Power
- Different learners have different power



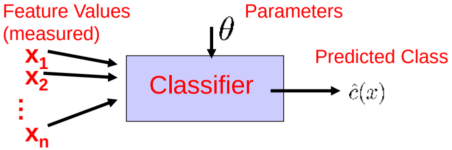


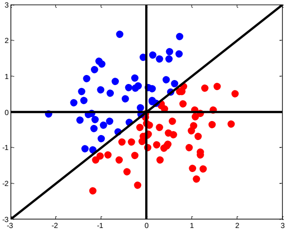
Example:

$$\hat{c}(x) = \text{sign}(\theta_1 x_1 + \theta_2 x_2 + \theta_0)$$

Learners and Complexity

- We've seen many versions of underfit/overfit trade-off
  - Complexity of the learner
  - Representational Power
- Different learners have different power



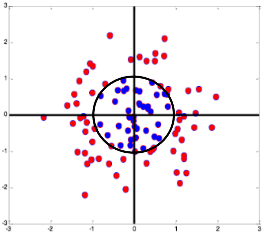


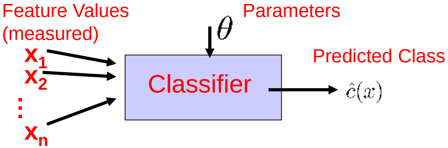
Example:

$$\hat{c}(x) = \text{sign}(\theta_1 x_1 + \theta_2 x_2)$$

Learners and Complexity

- We've seen many versions of underfit/overfit trade-off
  - Complexity of the learner
  - Representational Power
- Different learners have different power





Example:

$$\hat{c}(x) = \text{sign}(\,(x_1^2 + x_2^2) - \theta_0)$$

Learners and Complexity

- We've seen many versions of underfit/overfittrade-off
  - Complexity of the learner
  - Representational Power
- Different learners have different power
- Usual trade-off:
  - More power = represent more complex systems, might overfit
  - Less power = won't overfit, but may not find "best" learner
- How can we quantify representational power?
  - Not easily. . .
  - One solution is Vapnik-Chervonenkis (VC) dimension

Some notation

- Assume training data are iid from some distribution $p(X, Y)$
- Define "risk" and "empirical risk"
  - These are just "long term" test and observed training error

$$R(\theta) = \text{ Test Error } = \mathbb{E}[\mathbf{1}[c \neq \hat{c}(x;\theta)]]$$
$$R^{\text{emp}}(\theta) = \text{ Train Error } = \frac{1}{m}\sum_i \mathbf{1}\left[c^{(i)} \neq \hat{c}\left(x^{(i)};\theta\right)\right]$$

- How are these related? Depends on overfitting...
  - Underfitting domain: pretty similar...
  - Overfitting domain: test error might be lots worse!

VC Dimension and Risk

- Given some classifier, let H be its VC dimension represents "representational power" of classifier

$$R(\theta) = \text{ Test Error } = \mathbb{E}[\mathbf{1}[c \neq \hat{c}(x;\theta)]]$$
$$R^{\text{emp}}(\theta) = \text{ Train Error } = \frac{1}{n}\sum_i \mathbf{1}\left[c^{(i)} \neq \hat{c}\left(x^{(i)};\theta\right)\right]$$

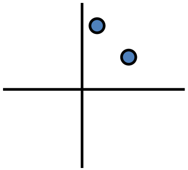
- With "high probability", Vapnik showed

$$\text{Test Error} \leq \text{Train Error} + \sqrt{\frac{H\log(2m/H) + H - \log(\eta/4)}{n}}$$

The bounds suggest that the optimism increases with h and decreases with n in qualitative agreement.

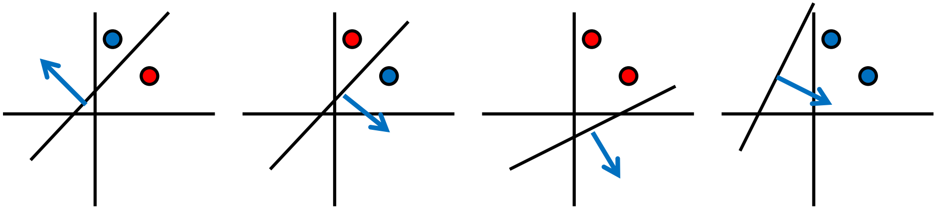
Shattering

- We say a classifier $f(x)$ can shatter points $x^{(1)} \dots x^{(h)}$ iff for all $y^{(1)} \dots y^{(h)}$, $f(x)$ can achieve zero error on training data $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \dots, \left(x^{(h)}, y^{(h)}\right)$ (i.e., there exists some θ that gets zero error)
- Can $f(x;\theta) = \operatorname{sign}\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2\right)$ shatter these points?



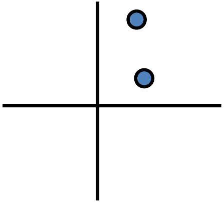
Shattering

- We say a classifier $f(x)$ can shatter points $x^{(1)} \ldots x^{(h)}$ iff for all $y^{(1)} \ldots y^{(h)}$, $f(x)$ can achieve zero error on training data $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(h)}, y^{(h)}\right)$ (i.e., there exists some θ that gets zero error)
- Can $f(x;\theta) = \operatorname{sign}\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2\right)$ shatter these points?
- Yes: there are 4 possible training sets...



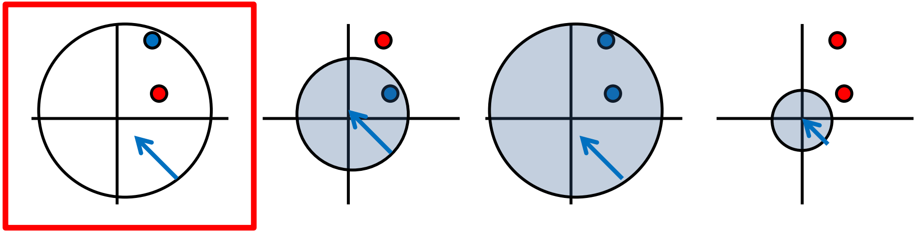
Shattering

- We say a classifier $f(x)$ can shatter points $x^{(1)} \dots x^{(h)}$ iff for all $y^{(1)} \dots y^{(h)}$, $f(x)$ can achieve zero error on training data $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \dots, \left(x^{(h)}, y^{(h)}\right)$ (i.e., there exists some θ that gets zero error)
- Can $f(x;\theta) = \operatorname{sign}\left(x_1^2 + x_2^2 - \theta\right)$ shatter these points?



Shattering

- We say a classifier $f(x)$ can shatter points $x^{(1)} \dots x^{(h)}$ iff for all $y^{(1)} \dots y^{(h)}$, $f(x)$ can achieve zero error on training data $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \dots, \left(x^{(h)}, y^{(h)}\right)$ (i.e., there exists some θ that gets zero error)
- Can $f(x;\theta) = \operatorname{sign}\left(x_1^2 + x_2^2 - \theta\right)$ shatter these points?
- Nope!

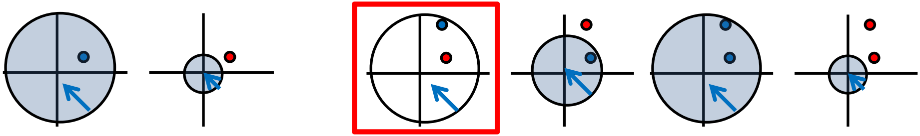


VC Dimension

- The VC dimension H is defined as "the maximum number of points h that can be arranged so that $f(x)$ can shatter them."
- The VC dimension of the class $\{f(x, \alpha)\}$ is defined to be the largest number of points (in some configuration) that can be shattered by members of $\{f(x, \alpha)\}$.
- A game:
  - Fix the definition of $f(x; \theta)$
  - Player 1: choose locations $x^{(1)} \dots x^{(h)}$
  - Player 2: choose target labels $y^{(1)} \dots y^{(h)}$
  - Player 1 : choose value of θ
  - If $f(x; \theta)$ can reproduce the target labels, $P1$ wins

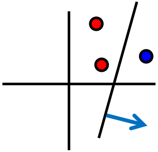
VC Dimension

- Example: what's the VC dimension of the (zero-centered) circle, $f(x;\theta) = \operatorname{sign}\left(x_1^2 + x_2^2 - \theta\right)$?
- VC dim $= 1$: can arrange one point, cannot arrange two (previous example was general)

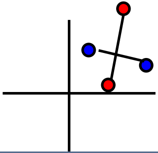


VC Dimension

- Example: what's the VC dimension of the two-dimensional line, $f(x;\theta) = \text{sign}\left(\theta_1 x_1 + \theta_2 x_2 + \theta_0\right)$?
- VC dim $= 3$? Yes



- VC dim $= 4$? No...



- Turns out: For a general, linear classifier (perceptron) in d dimensions with a constant term: VC dim $= d + 1$

VC Dimension

- VC dimension measures the "power" of the learner
- Does not necessarily equal the # of parameters!
- Number of parameters does not necessarily equal complexity
  - Can define a classifier with a lot of parameters but not much power (how?)
  - Can define a classifier with one parameter but lots of power (how?)
- Lots of work to determine what the VC dimension of various learners is...
- Vapnik's structural risk minimization (SRM) approach fits a nested sequence of models of increasing VC dimensions $h_1 < h_2 < \cdots$, and then chooses the model with the smallest value of the upper bound.

Using VC dimension

- Validation / cross-validation to select complexity
- VC dimension based bound on test error similarly
- Other Alternatives
  - Probabilistic models: likelihood under model (rather than classification error)
  - AIC (Akaike Information Criterion)
    - Log-likelihoood of training data - # of parameters

$$\text{AIC} = -\frac{2}{N} \cdot \log \text{lik} + 2 \cdot \frac{d}{N}$$

  - BIC (Bayesian Information Criterion)
    - Log-likelihood of training data - (# of parameters)$* \log(N)$

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

References

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Hastie, Tibshirani, and Friedman, Springer.
- In Introduction to Statistical Learning with Application in R, Second Edition, James, Witten, Hastie, and Tibshirani, Springer.

Thank you!