Lecture 17

Exam I (Thursday, Nov 2)

1) Logistic regression (Chap 24) — log odd ratio
2) Neural networks (Chap 25) — net gradients, D_C, D_b, output activations
3) Bayes classifier (Chap 18) — naive, full (no categorical data)
   → KNN classifiers
4) Decision Trees (Chap 19) — split eval, metrics
5) Support Vector Machines (Chap 21) —
   a) Margin vs. slacks
   b) Given $\xi^2$, compute $\vec{w}$ (linear)
   \[ \vec{w} = \sum_{i=1}^{n} a_i y_i \phi(x_i) \]
   c) Some iterations of SCA
      (find $\vec{w}$)
6) Classification assessment (Sec 22.1 only)
   ▶ Prec, recall, f1, roc (TPR vs FPR)

Classification Assessment

Sample $\rightarrow$ D $\rightarrow$ training set [80%]
validation set [20%] $\rightarrow$ testing set [20%]

Never look at this data
true response

User-specified constants

\(C\) hyperparameters

try diff C

values (grid search)

\[
\begin{bmatrix}
10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4, \ldots
\end{bmatrix}
\]

build model on training
curve on validation

Once we have a trained & validated model

\(\downarrow\) then apply on test data

\(\downarrow\) compute F1-score

---

**k-fold cross validation (CV)**

\(\downarrow\) repeat k times: build k different models

\(\downarrow\) compute \(\Theta_i\) \(i = 1 \ldots k\) (on the test data)

some measure, e.g., accuracy, F1

\[
\bar{\Theta} = \text{mean}\{\Theta_1, \Theta_2, \ldots, \Theta_k\}
\]

\[
\bar{\sigma}^2 = \text{var}\{\Theta_1, \Theta_2, \ldots, \Theta_k\}
\]

\(0.85\pm0.15\)
**k = s**

**s-fold CV**

1) **Shuffle D**
2) **Create s equal partitions**

3) **Keep aside Di for testing**
   We remaining folds for training & validation
   \[ T_i = D \setminus D_i \leftarrow \text{training} \]
   \[ D_i \leftarrow \text{testing} \]

\[ \text{LOOCV: leave one out CV} \]

Usualy used for data which is expensive to collect

\[ n - \text{fold CV} \]

1 point for testing \hspace{1cm} n-1 points for training

\[ \text{Sample} \]

\[ D \]

\[ \bar{D} \]

\[ \overbrace{D_1, D_2, D_3, \ldots, D_s}^{s - \text{folds}} \]

\[ D_1, D_2, D_3, D_4 \]

\[ D_0_i \leftarrow \text{model ( } D_i \text{) } \]
1 point for testing  \quad n-1 \text{ points for training}

Bootstrap sampling

Sample random dataset of size \(n\)

\[ D \quad \forall \quad i=1...k \]

Sample with replacement

Each point can be selected multiple times

Training set: \(D_i\) (further split into validation)

Test on \(D\)

\(D_i \subseteq D\)

\(\hat{O}_i \quad \text{over-estimate (optimistic estimate)}\)

---

Q: Sample with replacement, \(n\) trials

What is the probability that a point will not be sampled?

a) What is the probability that a point is sampled

\[ \frac{1}{n} \]

b) Prob that it is not sampled

\[ (1 - \frac{1}{n}) \text{ for one trial} \]

c) Not sampled even after \(n\) trials

\[ \sim \frac{1}{n} \]
\[
\mathbb{P}(y^*_j \neq D_i) = \left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0.3678
\]

A point is not sampled for \(D_i\):

\(D_i\) contains only about 65.2\% of the points!

From \(\hat{\mu} \pm \hat{\sigma}\) (mean & variance) for a classifier estimates we get a confidence bound on the true mean expected performance (accuracy).

\[\mu \sim \text{true mean, compute the confidence interval for the true mean} \]

\[
\left(\hat{\mu} - t_k(\hat{\sigma}) \leq \mu \leq \hat{\mu} + t_k(\hat{\sigma})\right)
\]

\(\ell_b \leq \mu \leq u_b\)

\(t_k\) : Student's t-distribution

\(\alpha\) : confidence level
\( \alpha : \text{confidence level} \) 90%.

\( d = \alpha \gamma \)

\( t_k \): student's t-distribution with \( k-1 \) degrees of freedom.

\( t_k \) is like the normal as \( k \to \infty \)

\( t_k \) is small sample version of normal

\( k \) to small e.g. \( k=5 \), \( k=10 \)

---

**How to compare two models?**

\( M_A \ & M_B \)

C(sum) (logistic regression)

\[ \hat{\mu}_A \pm \frac{\hat{\sigma}_A}{\sqrt{\hat{\gamma}}} \] vs \[ \hat{\mu}_B \pm \frac{\hat{\sigma}_B}{\sqrt{\hat{\gamma}}} \]

\( (\frac{1}{1}) \)

\( \mu_A \ & \mu_B \)

---

Paired t-test

**CV k-folds**

For \( i = 1 \ldots k \)

\[ T_i = D \setminus D^i \rightarrow \text{training} \]
\[
\begin{align*}
T_i &= D \setminus D_i \quad \text{training} \\
D_i &= i-i^{th} \text{ fold} \quad \text{testing} \\
\text{train } M_A \text{ & } M_B \text{ on } T_i \\
\Theta_i^A &= M_A(D_i) \\
\Theta_i^B &= M_B(D_i) \\
\hat{S}_i &= \Theta_i^A - \Theta_i^B \quad \text{difference in performance}
\end{align*}
\]

\[
\hat{m}_d = \text{mean difference} = \text{mean}\{S_1, S_2, \ldots, S_k\}
\]

\[
\hat{\sigma}_d^2 = \text{variance of difference}
\]

**Hypothesis testing**

- **\(H_0\)**: Are the two models different in terms of \(\Theta\) (accuracy)
  - **null hypothesis**
  - \(H_0\): Null hypothesis
    - (a) there is no difference between the two models
    - (b) look for evidence to support or reject \(H_0\)
  - **\(H_a\)**: alternative hypothesis
    - there is a difference

\[
\mu_8 = 0 \quad \text{true expected difference}
\]
The true expected difference is $\mu_8 = 0$.

### t-test

- **Hypothesis**
  - Mean $\mu_8$ = $\frac{\hat{\mu}_8}{\frac{S_8}{\sqrt{jk}}}$
  - Standardized difference: how many deviations away from the true mean

- **t-value**
  - $t_{a/2} < z_0 < +t_{a/2}$

- **Significance Level**
  - $\alpha = 0.01$
  - $\alpha = 0.05$

- **Decision**
  - If $Z_0$ lies outside the confidence interval then reject $H_0$.
  - If $Z_0$ lies in the confidence interval, $H_0$ is probably the right conclusion.

- **Confidence Interval**
  - This has a $t$-distribution with $k-1$ degrees of freedom.
Bias - Variance of the Classifier

Bias = Inherent class of functions a model can approximate.

\[
\text{bias} \quad \rightarrow \quad \text{high bias} \quad \rightarrow \quad \text{low variance}
\]

Linear models \quad \rightarrow \quad \text{high bias}

NN \quad \rightarrow \quad \text{low bias}

(Complex functions)

\[
\text{error} = \text{bias} + \text{variance}
\]

Bias - Variance tradeoff.