Lecture 8 Data Splits, Models & Cross Validation

1. bias - variance trade-off
   underfit : high bias
   overfit : high variance

2. Regularization
   examples: $\min_{\theta} \frac{1}{2} \sum_{i=1}^{m} \| y^{(i)} - \theta^T x^{(i)} \|^2 + \boxed{\lambda \|\theta\|^2}$

   $\operatorname{argmax}_{\theta} \sum_{i=1}^{n} \log P(y^{(i)} | x^{(i)}; \theta) - \boxed{\lambda \|\theta\|^2}$

   SVM: $\min \|w\|^2$ has the same effect

3. Bayesian statistics and regularization

training set $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$

$$p(\theta|S) = \frac{P(S|\theta)\cdot P(\theta)}{P(S)}$$

logistic regression

$$\arg\max_\theta P(\theta|S) = \arg\max_\theta P(S|\theta)P(\theta) = \arg\max_\theta \left(\prod_{i=1}^{m} P(y^{(i)}|x^{(i)},\theta)\right)\cdot P(\theta)$$

assume: $P(\theta) \sim N(0, \tau^2 I)$ (Gaussian)

$$P(\theta) = \frac{1}{\sqrt{2\pi}(\tau^2 I)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\theta^T(\tau^2 I)^{-1}\theta\right)$$

Frequentist:

$\arg\max_\theta P(S|\theta)$ - MLE

Bayesian:

prior distribution $P(\theta)$; $\arg\max_\theta P(\theta|S)$ - MAP estimation (maximum a posteriori)

4. Data Splits

S_{train}, S_{dev}, S_{test} (dev = development / cv data set)

Simple Hold-out Cross Validation:

1) Train each model i (option for degree of polynomial / λ / ...) on Strain. Get some hypothesis h_i.
2) Measure error on Sdev. Pick model with lowest error on Sdev.
3) Optional: apply algorithm to Stest and report the error

Traditionally: Strain 70% Sdev 30% or Strain 60% Sdev 20% Stest 20%
Now: if huge data set Strain 90% Sdev 5% Stest 5%

I. k-fold cross validation (small data set)
Leave-one-out CV (smaller data set, 20-50 rows)

6. Feature Selection
Start with $\mathcal{F} = \phi$
Repeat: {

1) Try adding each feature to F, and see which single feature addition most improve dev set performance
2) Add that feature to F

}