Model Selection Intro

Monday, November 12, 2018 10:59 PM

$h_\theta(x) = g(\theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_k x^k)$ ← k the degree

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

Ex.: SVM C $\mathcal{L}_1$ regularization

$\mathcal{M} = \{M_1, \dots, M_k\}$

Ex. $M_k = p_\theta^k$: $p_\theta^k(x) = \theta_0 + \theta_1 x + \dots + \theta_k x^k$

Ex. $k = 100$

Assume $|\mathcal{M}| < \infty$

Cross Validation

Monday, November 12, 2018 11:00 PM

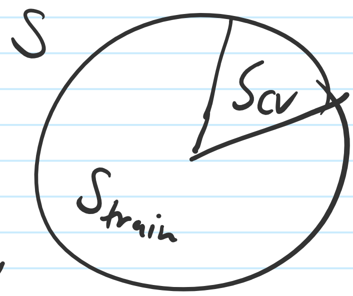
Empirical Risk: $\hat{\varepsilon}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}$

Version 1)

1. Train each M_i on $S \to h_i$
2. Pick h_j with smallest empirical error $\hat{\varepsilon}_S(h_i)$ $i = 1, \dots, n$

Bad idea
→ overfitting

Hold-out cross validation



1. Split S into S_{train} and S_{cv}
   $S_{train} \cap S_{cv} = \emptyset$ $S = S_{train} \cup S_{cv}$
2. Train each model M_i on S_{train}
   $\to h_i$
3. Select h_j that has smallest $\hat{\varepsilon}_{S_{cv}}(h_j)$

Test

Optionally: Retrain h_j on S

Typical fractions of S are $\frac{|S_{cv}|}{|S|} = 30\%$

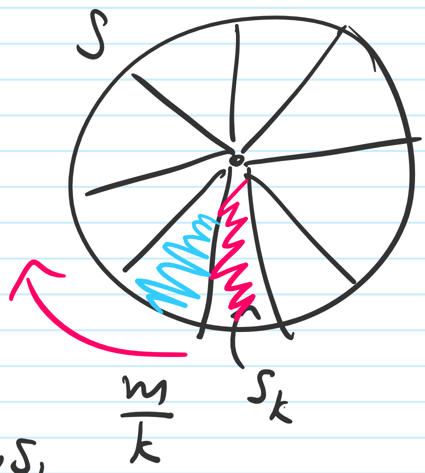
Disadvantage of hold-out CV ?

- If we don't have enough data, model selection is

- If we don't have enough data, model selection is poor.

---

k-fold cross validation



1. Randomly split S into $S_1, \dots, S_k$

2. For each model M_i

   For $j = 1, \dots, k$:

   Train M_i $S_1 \cup \dots S_{j-1} \cup S_{j+1} \dots \cup S_k$

   $\to h_{ij}$ evaluate $\hat{\varepsilon}_{S_j}(h_{ij})$

   $\hat{\varepsilon}_i = \sum_{j=1}^{k} \hat{\varepsilon}_{S_j}(h_{ij})$

3. Pick model M_j ith lowest $\hat{\varepsilon}_i$

   Optionally: Retrain M_i on S

Ex. $k=10$

⊕ More data to train

⊖ More computation needed

Leave-one-out cross validation: $k = m$

Feature Selection

Monday, November 12, 2018 11:00 PM

↳ Special case of model selection

$S = \{ (x^{(i)}, y^{(i)}) \mid i = 1, \ldots, m \}$ $\qquad x^{(i)} \in \mathbb{R}^n$

$n >> m$

$x^{(i)} \begin{pmatrix} \vdots \\ \vdots \\ \cancel{\vdots} \end{pmatrix} \in \mathbb{R}^n$ $\qquad \mathcal{M} = \{ M_i \}_{i=1,\ldots,2^n}$

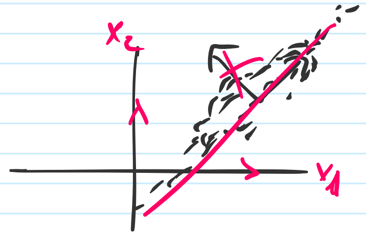
↳ Too many possibilities

Forward Search

1. Initialize $\mathcal{F} = \emptyset$, $\mathcal{F}_i = \emptyset$ $\quad i = 1, \ldots, n$
2. Repeat {
   a) For $i = 1, \ldots, n$ if $i \notin \mathcal{F}$ $\quad \mathcal{F}_i = \mathcal{F} \cup \{i\}$
   Use cross validation (simple, k-fold, hold-one-out)
   to evaluate $\mathcal{F}_i \leftrightarrow x'^{(i)} = \left(\vdots \right)$
   b) Set $\mathcal{F}$ to the best performing $\mathcal{F}_i$: $\mathcal{F} = \mathcal{F}_i$
   }
3. Select best feature subset $\mathcal{F}$ that was evaluated in the entire procedure $\quad \mathcal{F} \subseteq \{1, \ldots, n\}$

Terminate: if $|\mathcal{F}| > t$

Backward search: Initialize $\mathcal{F}$ as $\{1, \ldots, n\}$



Filter Feature Selection

- Heuristic
- Fast to compute

How well are x_i and y correlated? $\longrightarrow S(i)$

Sort $S(i)$ and select best x_i's.

Mutual information:

$$MI(\phi_i(x), y) = \sum_{\phi_i} \sum_{y} p(\phi_i, y) \log\left(\frac{p(\phi_i, y)}{p(\phi_i)\, p(y)}\right) =: S(i)$$

Ex.: $y \in \{0,1\}$

Ex: $\Phi_i(x) = x_i$

Compute score $S(i)$ for every $i = 1, \dots, n$

$$MI(\phi_i(x), y) = KL\Big(\overbrace{p(\phi_i(x), y)}^{\text{distr.}} \,\Big\|\, \overbrace{p(\phi_i(x))\, p(y)}^{\text{distr.}}\Big)$$

Kullback-Leibler divergence

(Compare two distributions)

Properties:
- $KL(p, q) \geq 0$
- $KL(p, q) = 0 \iff p = q$
- $KL(p, q) \neq KL(q, p)$

Bayesian Statistics and Regularization

Monday, November 12, 2018 11:01 PM

$$\boxed{\theta_{ML} = \arg\max_{\theta} \prod_{i=1}^{m} p_{\theta}(y^{(i)} \mid x^{(i)}; \theta)}$$

θ unknown ? $\qquad p(\theta)$

$$S = \{(x^{(i)}, y^{(i)})\}_{i=1,\dots,n}$$

$$p(\theta \mid S) = \frac{p(S,\theta)}{p(S)} \quad \circ \text{ (Bayes rule)}$$

$$= \frac{p(S,\theta)}{\int p(S,\theta)\, d\theta} \quad \circ \text{ Marginalization}$$

$$= \frac{\prod_{i=1}^{m} p(s_i \mid \theta)\, p(\theta)}{\int \left(\prod_{i=1}^{m} p(s_i \mid \theta)\, p(\theta)\right) d\theta} \quad \text{Bayes + indep.}$$

$$= \frac{\prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}, \theta)\, p(\theta)}{\int \left(\prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}, \theta)\, p(\theta)\right) d\theta} = p(\theta \mid S)$$

For new data y, x

$$p(y \mid x, S) = \int p(y, \theta \mid x, S)\, d\theta$$

Bayes ...

$$p(y|x,S) = \int p(y,\theta|x,S)\, d\theta$$

$$\overset{\text{Bayes}}{=} \int p(y|\theta,x,S)\, p(\theta|x,S)\, d\theta$$

$$= \int p(y|\theta,x)\, p(\theta|S)\, d\theta$$

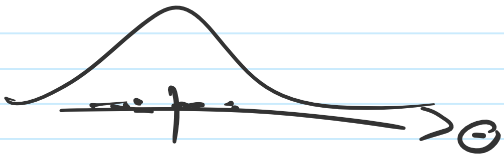
$$\boxed{\mathbb{E}(y|x,S) = \int y\, p(y|x,S)\, dy}$$

$p(\theta)$

The MAP (Maximum a posteriori) estimate of θ

$$\theta_{MAP} = \arg\max_{\theta} \prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta) \cdot p(\theta)$$

$$\theta_{ML} = \arg\max_{\theta} \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)$$



Ex. $p = \mathcal{N}(\mu_\theta, \Sigma_\theta)$ (for prior)

Ex. $p = \mathcal{N}(0, \sigma^2 I)$

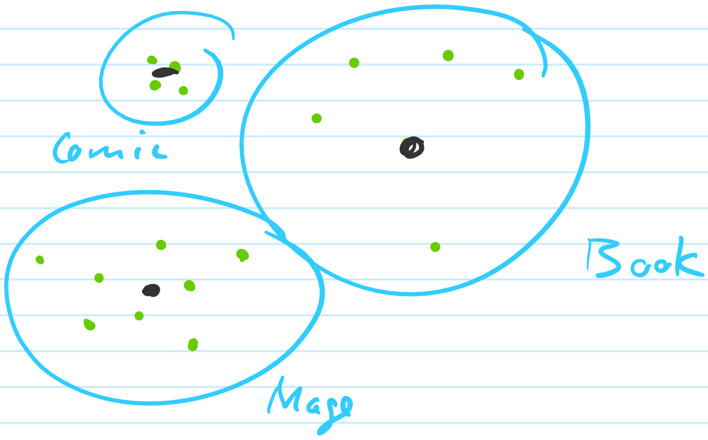
The k-means Clustering Algorithm

Monday, November 12, 2018 11:01 PM

- unsupervised → no labels $\boxed{y^{(i)}}$!

$S = \{x^{(i)}\}_{i=1,\dots,n}$

$x^{(i)} \in \mathbb{R}^n$

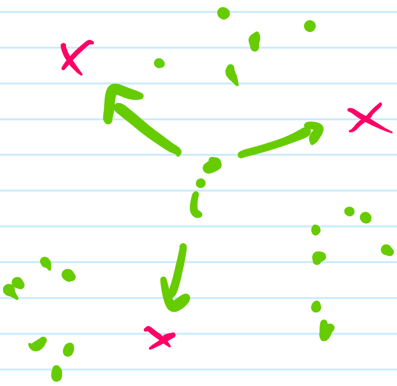


k-means k clusters

1. Initialize cluster centroids $\mu_1, \dots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: $\{$

   (a) For every $i = 1, \dots, m$

   $$c^{(i)} := \arg\min_j \|x^{(i)} - \mu_j\|^2$$

   (b) For each j we set

   $$\mu_j = \frac{\sum \mathbb{1}\{c^{(i)} = j\} x^{(i)}}{\sum \mathbb{1}\{c^{(i)} = j\}}$$

$\}$

Stop if $c^{(1)} \dots c^{(m)}$

Stop if $c^{(1)} \dots c^{(m)}$

don't change anymore