Regularization (Weight Decay)

Tuesday, February 5, 2019 11:32 AM

- Any Modification made to learning algorithm to reduce Test error, but not necessarily training error, is termed as regularization.

in fact, $E^{(train)}$ can increase.

- A particular approach to regularization is Weight decay.

$$\left\{ \min_{\underline{W}} \; J(\underline{W}) = \sum_{n=1}^{N} \left[y^{(n)} - t^{(n)} \right]^2 \quad \text{subject to} \quad \|\underline{W}\|_2^2 \le \rho \right\}$$

L_2-norm

equivalently

$$\min_{\underline{W}} \; J(\underline{W}) = \sum_{n=1}^{N} \left[y^{(n)} - t^{(n)} \right]^2 + \boxed{\lambda \|\underline{W}\|_2^2} \quad \text{for } \lambda \geq 0$$

L_2-norm → Ridge Regression

$L1$-norm → Lasso Regression, it may lead to more sparse solns.

$$\nabla_{\underline{W}} J(\underline{W}) = 2\underline{W}^T X^T X - 2t^T X + 2\lambda \underline{W}^T = 0$$

$$\boxed{(X^T X + \lambda I)\underline{W} = X^T \underline{t}}$$

Where I is $(D+1)\times(D+1)$ identity matrix

$$\boxed{\underline{W} = (X^T X + \lambda I)^{-1} X^T \underline{t}}$$

Regularization with nonlinear models

$$\min_{\underline{W}} \; J(\underline{W}) = \sum_{n=1}^{N} \left[\underline{W}^T \Phi(\underline{x}^{(n)}) - t^{(n)} \right]^2 + \lambda \underline{W}^T \underline{W}$$

$$\boxed{\underbrace{\underline{W}}_{(M+1)\times 1} = (\underbrace{\Phi^T \Phi}_{(M+1)\times(M+1)} + \underbrace{\lambda I}_{(M+1)\times(M+1)})^{-1} \Phi^T \underbrace{\underline{t}}_{N\times 1}}$$

Φ : $N\times(M+1)$

$(M+1) \times 1$

$(M+1) \times (M+1)$

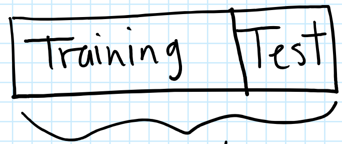
$(M+1) \times (M+1)$

Test Error Estimation & Hyperparameter Selection

Thursday, February 7, 2019 11:36 AM

1 Hold-out Method

→ randomize data set



→ 80% 20%

example or training set

- $\underline{W}$ is generated using training set & Test set is used with $\underline{W}$ to compute $E^{(test)}$

Problems

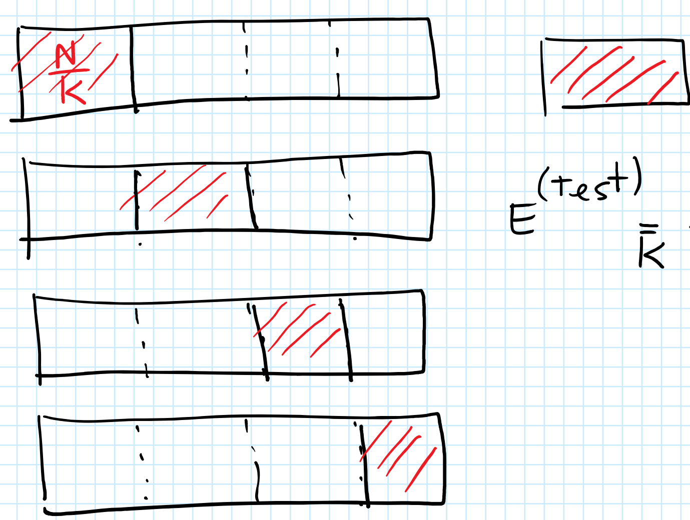
→ for small N (sparse datasets) this is not an efficient way to use data.

→ Unlucky & produce unfortunate splits

2 K-fold Cross-Validation

- K experiments, in each we use K-1 folds for training & remaining fold for testing

K=4



$$E^{(test)} = \frac{1}{K} \sum_{i=1}^{K} E_i^{(test)}$$

- All examples are eventually used for both Testing & training.

3 Leave-one-out is k-fold C.V. with k=N

Choice of K depends on N

- for sparse datasets, K is to be large
- for large N, $K=3$ is sufficient

& in general $3 \leq K \leq 10$

(4) Bootstrap: resampling technique w/ replacement

example: $N=5$ $\quad \underline{x}^{(1)} \quad \underline{x}^{(2)} \quad \underline{x}^{(3)} \quad \underline{x}^{(4)} \quad \underline{x}^{(5)}$

Exp. 1 $\quad \underbrace{\underline{x}^{(1)} \quad \underline{x}^{(3)} \quad \underline{x}^{(3)} \quad \underline{x}^{(3)} \quad \underline{x}^{(5)}}_{\text{Train. Set}} \qquad \underbrace{\underline{x}^{(2)} \quad \underline{x}^{(4)}}_{\text{test set}} \leftarrow$

Exp. 2 $\quad \underbrace{\underline{x}^{(1)} \quad \underline{x}^{(2)} \quad \underline{x}^{(3)} \quad \underline{x}^{(5)} \quad \underline{x}^{(5)}} \qquad \underbrace{\underline{x}^{(4)}}$

⋮

Exp. K

$$E^{(test)} = \frac{1}{K} \sum_{i=1}^{K} E_i^{(test)}$$

Hyperparameter Selection

Hyperparameter: Any setting (parameter) that is not adjusted during learning.

e.g., ρ, λ, M

1- Data is divided into 3 disjoint sets

Training Set	Validation Test	Test Set
60%	20%	20%

2- Select training parameters

3- train using training set

3 - train using training set

4 - Evaluate using Validation Set

5 - Repeat steps 2, 3 & 4

6 - Select the best Model & train with training & Validation Sets.

7 - Estimate test error, $E^{(test)}$, using test set.

Data Normalization

- Standardization (z-score)
- for each of D dimensions
    - subtract mean
    - divide by standard deviation
- Use this mean & standard deviation to normalize data from test set.

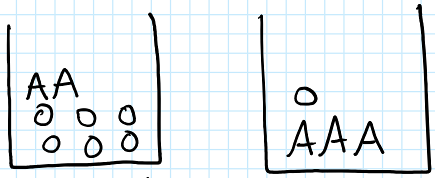
Probability (Review)

Thursday, February 7, 2019 11:37 AM

Red Box

A: Apples

O: oranges



40% 60%

Blue Box

$$P(B=r) = 0.4 \qquad P(B=b) = 0.6$$

$$\boxed{P(F=A) = ? \qquad P(B=r \mid F=O) = ?}$$

B & F are Random Variables

discrete

$P(F)$ & $P(B)$ are prob. mass functions

Three Rules: Sum Rule, Product Rule, Bayes Rule

$$P(F=A \mid B=r) = \frac{1}{4} \qquad P(F=O \mid B=r) = \frac{3}{4}$$

$$P(F=A \mid B=b) = \frac{3}{4} \qquad P(F=O \mid B=b) = \frac{1}{4}$$

joint prob. mass func. $\rightarrow$ $\underbrace{P(F, B)}_{\text{two events occurring simultaneously}} = P(F \mid B)\, P(B)$ $\leftarrow$ product rule

2-D Histogram

F \ B	r	b	
A	$\frac{1}{10}$	$\frac{9}{20}$	$P(F=A) = \frac{11}{20}$
O	$\frac{3}{10}$	$\frac{3}{20}$	$P(F=O) = \frac{9}{20}$
	$P(B=r) = 4/10$	$P(B=b) = 6/10$	

$\leftarrow$ Sum Rule

$$P(F=A, B=r) = P(F=A \mid B=r) \cdot P(B=r) = \frac{1}{4} \cdot \frac{4}{10} = \frac{1}{10}$$

$$P(F) = \boxed{\sum_{B} P(F, B)}$$

$$\underbrace{P(F=A)}_{\text{marginal prob.}} = P(F=A, B=r) + P(F=A, B=b) = \frac{1}{10} + \frac{9}{20} = \frac{11}{20}$$

marginalization

Posteriori $\rightarrow$ $$P(B|F) = \frac{\boxed{P(F|B)}\,\boxed{P(B)}}{\boxed{P(F)}} \leftarrow \text{Bayes Rule}$$

likelihood, prior

marginal likelihood

- Converts priors to posteriors
exploits role of evidence

$$P(B=r|F=o) = \frac{P(F=o \mid B=r)\,P(B=r)}{P(F=o)} = \frac{\frac{3}{4}\cdot\frac{4}{10}}{\frac{9}{20}} = \frac{2}{3}$$

without evidence $\quad P(B=r) = \frac{4}{10} = 0.4$

with " $\quad P(B=r|F=o) = 0.67$

Statistical Independence between R.V. if

$$\underline{P(F,B) = P(F)\cdot P(B)}$$

or $$\underline{P(F|B) = P(F)}$$

---

for Continuous R.V.

Instead of Prob. Mass functions, we have " density ".

$p(x)$ prob. density func. for a R.V. X

$$\text{prob.}\left(x \in (a,b)\right) = \int_a^b p(x)\,dx$$

$$p(x) \geq 0$$
$$\int_{-\infty}^{\infty} p(x) = 1$$

Say we have two R.V. X & Y

Sum Rule $\rightarrow$ $p(x) = \int p(x,y)\,dy$ where $p(x,y)$ joint prob. density func.

product Rule $\longrightarrow$ $p(x,y) = p(x|y)\,p(y)$

$= p(y|x)\,p(x)$

Expectation operator $\longrightarrow$ $E(\underline{x}) = \int x\, p(x)\, dx$ $\qquad E[f(\underline{x})] = \int f(x)\, p(x)\, dx$

Conditional expectation $\quad E(x|y) = \int x\, p(x|y)\, dx \;\leftarrow$ func. of y

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}$$

S.I. $\Leftrightarrow$ $p(x,y) = p(x)\cdot p(y)$

Y \ X	129	130	131
15	0.12	0.42	0.06
16	0.08	0.28	0.04

$P(X=130 \mid T=15) = ?$

$E(x) = ?$

$E(T \mid X=129) = ?$

Are X & T S.I. ?

Maximum Likelihood Estimate

Tuesday, February 12, 2019 12:08 PM

Given a set of samples $\{ \underline{x}^{(1)}, \underline{x}^{(2)} \ldots \underline{x}^{(N)} \}$ we would like to estimate the joint prob. density (pdf) $p(\underline{x}^{(1)}, \underline{x}^{(2)}, \ldots, \underline{x}^{(N)})$, assuming we know the form of pdf & attempt to estimate its parameters, $\underline{\underline{\theta}}$.

$\underline{x}^{(1)} \quad \underline{x}^{(2)} \quad \ldots \quad \underline{x}^{(N)}$
they came from a Normal distribution } Given
What are μ & σ ?

$$\underline{\theta} = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$$

$\longrightarrow$ Assume $\underline{x}$'s are i.i.d.

$$p(\underline{x}^{(1)}, \underline{x}^{(2)}, \ldots \underline{x}^{(N)}; \underline{\theta}) = \prod_{i=1}^{N} p(\underline{x}^{(i)}; \underline{\theta})$$

this function is called the $\boxed{\text{likelihood function}}$ of $\underline{\theta}$ with respect to examples & MLE finds $\underline{\theta}$ for which this function is maximum, i.e.,

$$\underline{\theta}_{ML} = \max_{\underline{\theta}} \prod_{i=1}^{N} p(\underline{x}^{(i)}; \underline{\theta})$$

equivalently

$$\underline{\theta}_{ML} = \max_{\underline{\theta}} \log \prod_{i=1}^{N} p(\underline{x}^{(i)}; \underline{\theta})$$

log is a monotonically increasing func.

$$x_1 \geq x_2 \qquad \log x_1 \geq \log x_2$$

$x_1 \geq x_2 \qquad \log x_1 \geq \log x_2$

$$\boxed{\underline{\theta}_{ML} = \max_{\underline{\theta}} \sum_{i=1}^{N} \underbrace{\log P(\underline{x}^{(i)} ; \underline{\theta})}_{\text{log-likelihood func.}}}$$