CSE 417T Machine Learning

Hw4, Due: Oct.27, 2016

Chih Yun Pai

Problem 4.4

(a) Why do we normalize f? In addition to answering the question about why we need to normalize f, also prove that the term to normalize by is $\sqrt{\sum_{q=0}^{Q_f} \frac{1}{2q+1}}$

$\rightarrow E_{a,x}[f^2] = E_{a,x}\left[(a_0L_0(x) + a_1L_1(x) + \cdots + a_{Q_f}L_{Q_f}(x))^2\right]$

$\because \int_{-1}^{1} L_k(x)L_l(x)\,dx = \begin{cases} 0, l \neq k \\ \frac{2}{2k+1}, l = k \end{cases}$ and $E_x[f(x)] = \int_a^b P(x)f(x)dx$

$\Rightarrow E_{a,x}[f^2] = E_a\left[\int_{-1}^{1} \frac{1}{2}[(a_0L_0(x))^2 + (a_1L_1(x))^2 + \cdots + (a_{Q_f}L_{Q_f}(x))^2]dx\right]$

$= E_a\left[\sum_{q=0}^{Q_f} \frac{a_q^2}{2q+1}\right]$

$= \frac{1}{1}E[a_0^2] + \frac{1}{3}E[a_1^2] + \frac{1}{5}E[a_2^2] + \cdots + \frac{1}{2Q_f+1}E[a_{Q_f}^2]$

$\because \sigma^2 = 1, then\ E[a_i^2] = \sigma^2 - (E[a_i^2])^2 = 1 - 0 = 1, \text{for } i = 1,2,\dots,Q_f$

$\Rightarrow \frac{1}{1}E[a_0^2] + \frac{1}{3}E[a_1^2] + \frac{1}{5}E[a_2^2] + \cdots + \frac{1}{2Q_f+1}E\left[a_{Q_f}^2\right] = \frac{1}{1} + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2Q_f+1} = \sum_{q=0}^{Q_f} \frac{1}{2q+1}$

$\Rightarrow E_{a,x}[f^2] = \sum_{q=0}^{Q_f} \frac{1}{2q+1}$

Letting this result to 1, the term can be normalized when each a_i, (for $i = 1,2,\dots,Q_f$) is divided by the normalizer $\sqrt{\sum_{q=0}^{Q_f} \frac{1}{2q+1}}$, that is, $\tilde{a}_i = \frac{a_i}{\sqrt{\sum_{q=0}^{Q_f} \frac{1}{2q+1}}}$.

The reason to normalize f is that normalizing $E_{a,x}[f^2] = 1$ let the noise level σ^2 automatically calibrated to the signal level.

(b) How can we obtain g_2 and g_{10}?

Our target function is $f(x) = \sum_{q=0}^{Q_f} a_qL_q(x)$ and hypothesis is $g_d(x) = \sum_{i=0}^{d} w_iL_i(x)$. Therefore, 2-order and 10-order polynomial model are $g_2(x) = \sum_{i=0}^{2} w_iL_i(x) = w^TL,\ w, L \in R^3$ and $g_{10}(x) = \sum_{i=0}^{10} w_iL_i(x) = w^TL,\ w, L \in R^{11}$

In implementation,

```
z_train_2 = L(x_train,2);
z_train_10 = L(x_train,10);
w_2  =  glmfit(z_train_2, y_train, 'normal', 'constant', 'off')
w_10  =  glmfit(z_train_10, y_train, 'normal', 'constant', 'off')
g_2  =  computeLegPoly( x_test, 2 ) * w_2
g_10  =  computeLegPoly( x_test, 10 ) * w_10
```

(c) How can we compute E_{out} analytically for a given g_{10}?

$E_{out} = \ E_{x,\epsilon}[(g_{10}(x) - y)^2]$
$= \ E_{x,\varepsilon}[(g_{10}(x) - (f(x) + \sigma\epsilon))^2]$
$= \ E_{x,\varepsilon}[((g_{10}(x) - f(x)) - \sigma\epsilon)^2]$
$= \ E_{x,\varepsilon}[\left(g_{10}(x) - f(x)\right)^2 - 2\left(g_{10}(x) - f(x)\right)\sigma\epsilon + (\sigma\epsilon)^2\Big]$

$= \ E_{x,\varepsilon}[\ (\sum_{i=0}^{10} w_i L_i(x) + \sum_{j=0}^{Q_f} a_j\, L_j(x))^2 - 2(\sum_{i=0}^{10} w_i L_i(x) - \sum_{j=0}^{Q_f} a_j\, L_j(x))\sigma\epsilon + (\sigma\epsilon)^2]$

$\because E_\varepsilon[\varepsilon] = 0 \quad \Rightarrow E_{x,\varepsilon}\left[-2\left(\sum_{i=0}^{10} w_i L_i(x) - \sum_{j=0}^{Q_f} a_j\, L_j(x)\right)\sigma\epsilon\right] = 0$

$\because E_\varepsilon[\varepsilon^2] = 1 \quad \Rightarrow E_{x,\varepsilon}[(\sigma\epsilon)^2] = \ \sigma^2$

$\Rightarrow E_{out} = \ E_{x,\varepsilon}[\ (\sum_{i=0}^{10} w_i L_i(x) + \sum_{j=0}^{Q_f} a_j\, L_j(x))^2]$

$= \ E_{x,\varepsilon}[\ (\sum_{i=0}^{10} w_i L_i(x))^2 - 2\sum_{j=0}^{\min(10,Q_f)}(w_j a_j\, L_j(x))^2 + \sum_{k=0}^{Q_f}(a_k\, L_k(x))^2]$

$= \ E_\varepsilon[\ \sum_{i=0}^{10}\frac{w_i^2}{2i+1} - 2\sum_{j=0}^{\min(10,Q_f)}\frac{w_j a_j}{2j+1} + \sum_{k=0}^{Q_f}\frac{a_k^2}{2k+1}]$

$= \ \sum_{i=0}^{10}\frac{w_i^2}{2i+1} - 2\sum_{j=0}^{\min(10,Q_f)}\frac{w_j a_j}{2j+1} + \sum_{k=0}^{Q_f}\frac{a_k^2}{2k+1}$

(d) Vary Q_f, N, σ, where $Q_f \in$ {5, 10, 15, 20}, N ∈ {40, 80, 120}, σ2 ∈ {0, 0.5, 1.0, 1.5, 2.0}, run 500 of experiments for each combination of the parameters, each time computing $E_{out}(g_2)$ and $E_{out}(g_{10})$. Averaging these out-of-sample errors gives estimates of the expected out-of-sample error for the given learning scenario (Q_f, N, σ) using H_2 and H_{10}. Define the overfit measure $E_{out}(H_{10})$ - $E_{out}(H_2)$.

$$\begin{cases} E_{out}(H_2) & = \text{average over experiments } (E_{out}(g_2)) \\ E_{out}(H_{10}) & = \text{average over experiments } (E_{out}(g_{10})) \end{cases}$$

· When is the overfit measure significantly positive (i.e., the overfitting is serious) as opposed to significantly negative?

→ Serious overfitting happens when using mean error measurement, model complexity Q_f is large, number of training data N is small and noise coefficient σ is large.

· How error measurement varies? Explain your observation.

→ In general, error measurement grows (negatively) larger when model complexity Q_f and N goes larger.

→ Given a sufficiently large N (N=80 or 120), the difference between complex model (H_{10}) and simple model (H_2) grows larger when model complexity Q_f grows larger.

→ Without stochastic noise, error measurement is larger than those in the same N and Qf but with stochastic noise.

→ In general,

overfitting goes up when stochastic noise σ^2 goes up. (more obvious when N is small)

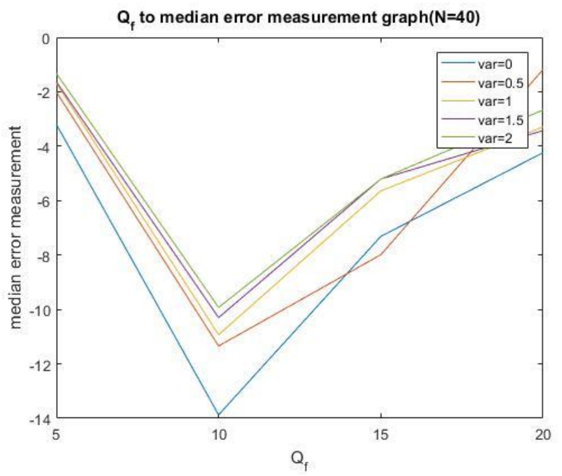
overfitting goes up when model complexity Q_f goes up.

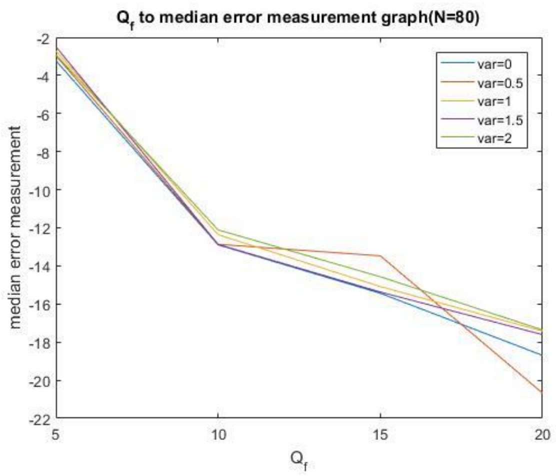
overfitting goes down when number of training data N goes up.

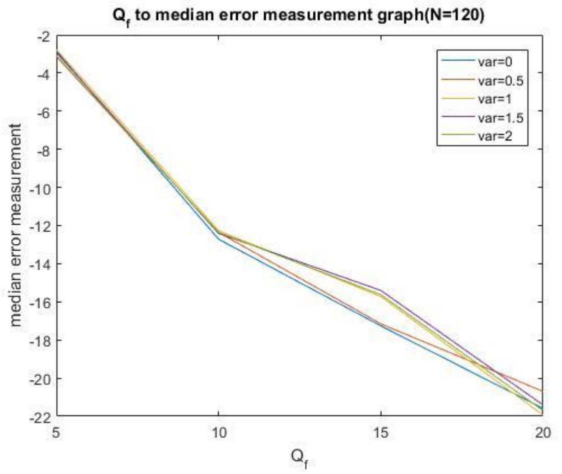
Comments on your observation on the error measurement between mean and median.

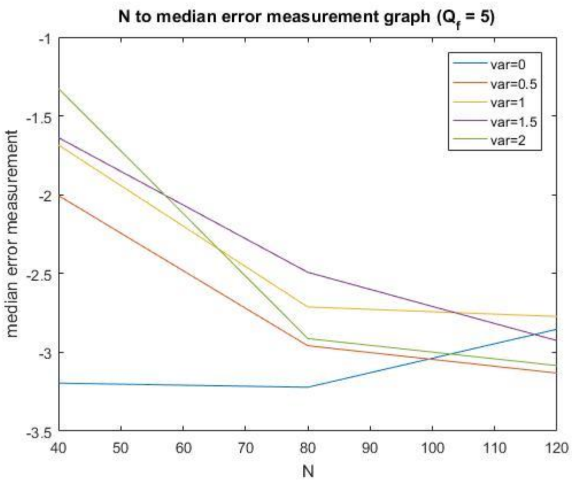
→ The result using mean error measurement goes very stable, because it counts very extreme case into average (e.g. very extreme positive or negative large ε or σ .

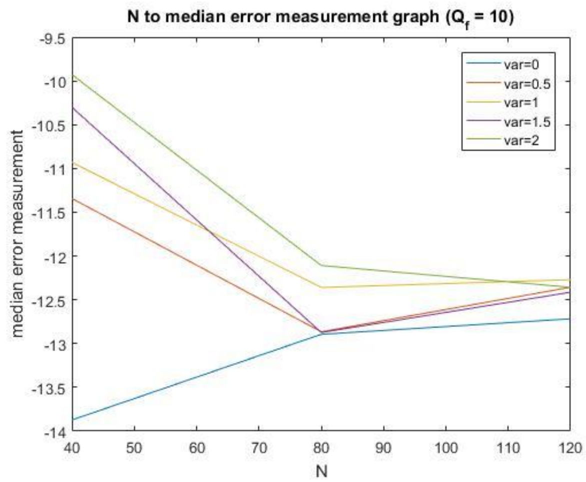
→ Using median's result relatively consistent due to picking median value of the error measurement vector.

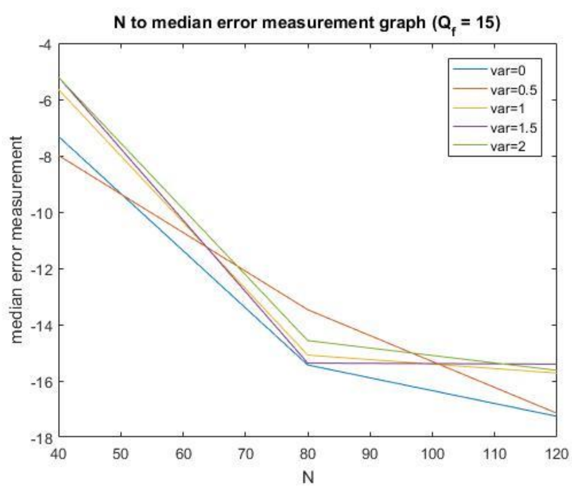

Q_f to median error measurement graph(N=40)
median error measurement
Q_f
var=0
var=0.5
var=1
var=1.5
var=2

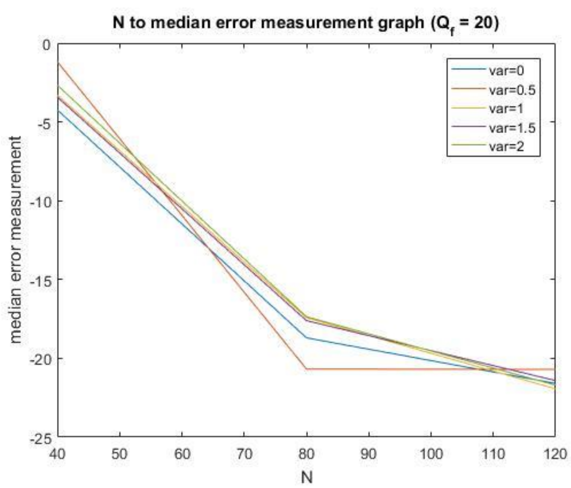

Q_f to median error measurement graph(N=80)
median error measurement
Q_f
var=0
var=0.5
var=1
var=1.5
var=2


Q_f to median error measurement graph(N=120)
median error measurement
Q_f
var=0
var=0.5
var=1
var=1.5
var=2


N to median error measurement graph (Q_f = 5)
median error measurement
N
var=0
var=0.5
var=1
var=1.5
var=2


N to median error measurement graph (Q_f = 10)
median error measurement
N
var=0
var=0.5
var=1
var=1.5
var=2


N to median error measurement graph (Q_f = 15)
median error measurement
N
var=0
var=0.5
var=1
var=1.5
var=2


N to median error measurement graph (Q_f = 20)
median error measurement
N
var=0
var=0.5
var=1
var=1.5
var=2

Result data (mean / median error measurement)

Mean error measurement | Median error measurement

var = 0	N = 40	N = 80	N = 120	var = 0	N = 40	N = 80	N = 120
Qf = 5	-4.12172	-4.01457	-3.86343	Qf = 5	-3.19543	-3.22157	-2.85317
Qf = 10	-15.1925	-14.774	-13.9225	Qf = 10	-13.8729	-12.8938	-12.7172
Qf = 15	3338.675	-6.90773	-17.238	Qf = 15	-7.31356	-15.4362	-17.2614
Qf = 20	2845.974	0.950951	-23.483	Qf = 20	-4.23798	-18.6988	-21.5905
var = 0.5	N = 40	N = 80	N = 120	var = 0.5	N = 40	N = 80	N = 120
Qf = 5	48.0182	-3.74689	-4.00618	Qf = 5	-2.00527	-2.95798	-3.13175
Qf = 10	57.57465	-14.2847	-13.5448	Qf = 10	-11.343	-12.8651	-12.3565
Qf = 15	3565.225	-5.04213	-14.4552	Qf = 15	-7.98864	-13.4717	-17.1567
Qf = 20	1663.051	-0.5797	-18.165	Qf = 20	-1.18132	-20.68	-20.7001
var = 1.0	N = 40	N = 80	N = 120	var = 1.0	N = 40	N = 80	N = 120
Qf = 5	209.6775	-2.94872	-3.66224	Qf = 5	-1.68482	-2.71218	-2.77246
Qf = 10	15.25814	-12.569	-13.6326	Qf = 10	-10.9303	-12.3592	-12.2701
Qf = 15	340.7107	-0.81053	-10.6132	Qf = 15	-5.64385	-15.0851	-15.7239
Qf = 20	2108.379	-8.60978	-21.6239	Qf = 20	-3.29368	-17.4243	-21.9367
var = 1.5	N = 40	N = 80	N = 120	var = 1.5	N = 40	N = 80	N = 120
Qf = 5	79.00667	-2.78508	-3.65293	Qf = 5	-1.6375	-2.49205	-2.92583
Qf = 10	49.38114	-14.3603	-14.0952	Qf = 10	-10.3007	-12.8735	-12.4134
Qf = 15	870.3479	0.403971	-16.1919	Qf = 15	-5.20563	-15.3705	-15.4035
Qf = 20	7970.857	3.369453	-21.0473	Qf = 20	-3.42507	-17.6101	-21.4139
var = 2.0	N = 40	N = 80	N = 120	var = 2.0	N = 40	N = 80	N = 120
Qf = 5	1157.489	-3.75437	-3.69793	Qf = 5	-1.32536	-2.91244	-3.08431
Qf = 10	279.0018	-12.8372	-13.5044	Qf = 10	-9.92836	-12.1084	-12.3572
Qf = 15	98481.64	-4.21175	-16.3296	Qf = 15	-5.20891	-14.5705	-15.6274
Qf = 20	7227.035	8.334017	-20.2116	Qf = 20	-2.67036	-17.3606	-21.7036