COMS 4771 Machine Learning (Spring 2015)
Problem Set #3

Jingwei Yang - `jy2653@columbia.edu`
Discussants: `pp2526,yd2300`

March 26, 2015

Problem 1

In order filter out the optimal parameters for the classifier, I have validated 5 * 11 classifiers with different values of λ and h on the training set.

In the process of selecting optimal parameters, I have used hold-out validation over each classifier. For each validation, I have selected 70% training data as training set and the remained 30% as validation set. The training set and validation set were randomly chosen from the overall training data. In order to achieve the reproducibility and debugging, I follow the advice of Professor Hsu to fix the random seed for each validation.
The selecting process could be depicted by following pseudocode.

Algorithm 1 Finding optimal parameters

$optAccuracy = 0$
$seed = rng$
for each λ in $lambdaList$ **do**
  for each $hRate$ in $hRateList$ **do**
    [$tempAccuracy$, h] = getAccuracy($data$, $labels$, λ, $hRate$, $seed$)
    if $tempAccuracy > optAccuracy$ **then**
      $optAccuracy = tempAccuracy$
      $optLambda = \lambda$
      $optH = h$;
    end if
  end for
end for
return $optLambda, optH$

Below, I list the pairs of parameters and related validation accuracy. According to the result, when $\lambda = exp(-9)$ and $h = 33.54$, the resulting classifier could have better performance over other classifiers.

Table 1: Hold-out validation(accuracy) for different classifiers with respective λ and h

λ \ h	18.98	33.54	63.03	115.75	219.43
1	0.6267	0.5865	0.5778	0.5756	0.5756
EXP(-1)	0.6267	0.5865	0.5778	0.5756	0.5756
EXP(-2)	0.6322	0.5919	0.5778	0.5767	0.5756
EXP(-3)	0.7595	0.7159	0.6256	0.5843	0.5756
EXP(-4)	0.8879	0.8639	0.8248	0.741	0.6191
EXP(-5)	0.9107	0.9042	0.8792	0.8498	0.8073
EXP(-6)	0.9173	0.914	0.9064	0.889	0.8607
EXP(-7)	0.9325	0.9249	0.914	0.9075	0.8966
EXP(-8)	0.9357	0.9314	0.9096	0.9075	0.9042
EXP(-9)	0.9336	0.9357	0.9129	0.9053	0.8955
EXP(-10)	0.9216	0.9281	0.9053	0.8998	0.8781

Using $\lambda = exp(-9)$ and $h = 33.54$, I test the classifier on the test data, and have achieved the following result.

Table 2: Test result

	Predict -1	Predict +1
Label -1	0.5846	0.0280
Label +1	0.0358	0.3516

Problem 2

(a)

Since we have

$$min_{\boldsymbol{w}\in R^d}\frac{\lambda}{2}||\boldsymbol{w}||_2^2+\frac{1}{|S|}\sum_{(\boldsymbol{x},y)\in S}(\langle\boldsymbol{w},\boldsymbol{x}\rangle-y)^2$$

and the function $f(\boldsymbol{w})$ is twice-differentiable,

$$\nabla f(\boldsymbol{w}) := \lambda\boldsymbol{w}+\frac{1}{|S|}\sum_{(\boldsymbol{x},y)\in S}2(\langle\boldsymbol{w},\boldsymbol{x}\rangle-y)\boldsymbol{x}$$
$$\nabla^2 f(\boldsymbol{w}) := \lambda\mathbb{I}+\frac{2}{|S|}\sum_{(\boldsymbol{x},y)\in S}\boldsymbol{x}\boldsymbol{x}^T$$

and we have

$$\langle\nabla^2 f(\boldsymbol{w})\boldsymbol{v},\boldsymbol{v}\rangle = \lambda\langle\boldsymbol{v},\boldsymbol{v}\rangle+\frac{2}{|S|}\sum_{(\boldsymbol{x},y)\in S}\langle\boldsymbol{v},\boldsymbol{x}\rangle^2\geq 0$$

Thus the optimization problem is convex.

(b)
From the previous inference, we have:

$$\nabla f(\boldsymbol{w}^{(t)}) = \lambda\boldsymbol{w}+\frac{1}{|S|}\sum_{(\boldsymbol{x},y)\in S}2(\langle\boldsymbol{w},\boldsymbol{x}\rangle-y)\boldsymbol{x}$$

Algorithm 2 The algorithm for solving the optimization problem

Start with some initial $\boldsymbol{w}_{(1)}\in\mathbb{R}^d$
for t = 1, 2, ... until some stopping condition is satisfied **do**
  Compute gradient of f at $\boldsymbol{w}^{(t)}$:
    $\boldsymbol{\lambda}^t := \nabla f(\boldsymbol{w}^{(t)})$
  Update:
    $\boldsymbol{w}^{(t+1)} := \boldsymbol{w}^{(t)} - \eta_t\boldsymbol{\lambda}^{(t)}$;
end for
return $\boldsymbol{w}$

(c)
The optimization problem is still convex. We can write the constraints $f(w_i)$ in the standard form

$$f(w_i) = w_i^2 - 1 \leq 0$$

Since $f(w_i)$ is twice-differentiable, we have

$$\nabla f(w_i) = 2w_i$$
$$\nabla^2 f(w_i) = 2 > 0$$

The $f(w_i)$ is convex, thus the optimization problem is still a convex optimization problem.

(d)

The optimization problem is still convex. We can write the constraints $f(w_i)$ in the standard form

$$f_1(w_i, w_{i+1}) = w_i - w_{i+1} \leq 0$$
$$f_2(w_i, w_{i+1}) = w_{i+1} - w_i \leq 0$$

for $f_1(w_i, w_{i+1})$, we have

$$\nabla f_1(w_i, w_{i+1}) = \begin{bmatrix} 1 & -1 \end{bmatrix}$$
$$\nabla^2 f_1(w_i, w_{i+1}) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \geq \mathbf{0}$$

for $f_2(w_i, w_{i+1})$, we have

$$\nabla f_2(w_i, w_{i+1}) = \begin{bmatrix} -1 & 0 \end{bmatrix}$$
$$\nabla^2 f_1(w_i, w_{i+1}) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \geq \mathbf{0}$$

Therefore, the problem is still a convex optimization problem.

(e)

The optimization problem is not convex. We can write the constraints $f(w_i)$ in the standard form

$$f_1(w_i) = w_i^2 - 1 \leq 0$$
$$f_2(w_i) = 1 - w_i^2 \leq 0$$

for $f_1(w_i)$, we have

$$\nabla f_1(w_i) = 2w_i$$
$$\nabla^2 f_1(w_i) = 2 > 0$$

for $f_2(w_i)$, we have

$$\nabla f_2(w_i) = -2w_i$$
$$\nabla^2 f_2(w_i) = -2 < 0$$

Since the constraints is not convex, the optimization problem is not convex.

Problem 3

(a)

Suppose $(\boldsymbol{x}, y)$ is randomly picked from the distribution P, and $err(f^*)$ is the probability that classifier $err(f^*)$ would wrongly classify $(\boldsymbol{x}, y)$. Since the size of the sample is $|A|$, the algorithm must meet the condition $Prb(err(f^*, A) = 0)$ to return the classifier. Thus, we could have following inference:

$$Prb(err(f^*, A) = 0) = (1 - err(f^*))^{|A|} \leq e^{-err(f^*)|A|} \leq e^{-\epsilon|A|}$$

As proved above, the bound is decrease exponentially with $|A|$.

(b)

Given

$$|S| > \frac{ln(|F|/\delta)}{\epsilon}$$

To prove $Prb[\forall f \in \mathcal{F} : err(f, |S|) = 0 \Rightarrow err(f) \leq \epsilon] \geq 1 - \delta$ is equal to prove $Prb[\forall f \in \mathcal{F} : err(f, |A|) = 0 \Rightarrow err(f) > \epsilon] < \delta$. Thus we could have following inference.

$$\begin{aligned} Prb[\forall f \in \mathcal{F} : err(f, |S|) = 0 \Rightarrow err(f) > \epsilon] &= Prb[\forall f \in \mathcal{F} : err(f, |S|) = 0 \wedge err(f) > \epsilon] \\ &\leq \sum_{i=1}^{|\mathcal{F}|} e^{-\epsilon|S|} \\ &\leq |\mathcal{F}| e^{-\epsilon|S|} \end{aligned}$$

Since we have $|S| > \frac{ln(|\mathcal{F}|/\delta)}{\epsilon}$, we could have

$$Prb[\forall f \in \mathcal{F} : err(f, |A|) = 0 \Rightarrow err(f) > \epsilon] \leq |\mathcal{F}| e^{-\epsilon|S|} < |\mathcal{F}| e^{-\epsilon \frac{ln(|\mathcal{F}|/\delta)}{\epsilon}} = \delta$$

Thus, we have proved desired claim through its complementary case.

(c)

Since we have training examples from different but independent distributions, the algorithm could return a classifier when $Prb(err(f, S)) = 0$.

$$Prb(err(f, S) = 0) = \prod_{i=1}^{n} (1 - err_i(f)) \leq \prod_{i=1}^{n} e^{-err_i(f)} = e^{\sum_{i=1}^{n} -err_i(f)}$$

Since P is the uniform mixture of $P_1, P_2, ..., P_n$, we also have

$$err(f) = \frac{1}{n} \sum_{i=1}^{n} err_i(f)$$

Thus we could have

$$Prb(err(f,S)=0) \leq e^{-err(f)|S|}$$

Then we could use the same inference route in part b to prove the the claim.
(d)

Since the overall failure rate is δ and $\sum_{t=1}^{\infty} \frac{1}{t(t+1)} \leq 1$, rather than split the δ equally among the the classifier among $f_1, f_2, f_3, ...$, we assign classifier f_t with the error bound

$$\delta_t = \frac{\delta}{t(t+1)}$$

Then we can mimic the inference on the 11_{th} page of slides 12 : Introduction to Learning theory. Apply to events ϵ_f for $f \in \mathcal{F}$ given by

$$\epsilon_f = \{err(f) > \mathrm{err}(f,S) + \sqrt{\frac{2\mathrm{err}(f,S)\ln(t(t+1)/\delta)}{|S|}} + \frac{2\ln(t(t+1)/\delta)}{|S|}\}$$

Therefore, use union bound, we could have

$$Prb[\forall f \in \mathcal{F}.err(f) \leq \mathrm{err}(f,S) + \sqrt{\frac{2\mathrm{err}(f,S)\ln(t(t+1)/\delta)}{|S|}} + \frac{2\ln(t(t+1)/\delta)}{|S|}] \geq 1-\delta$$

Since the Consistent Classifier Algorithm return $\hat{f} \in \mathcal{F}$, we know that

$$Prb[err(\hat{f}) \leq \mathrm{err}(\hat{f},S) + \sqrt{\frac{2\mathrm{err}(\hat{f},S)\ln(t(t+1)/\delta)}{|S|}} + \frac{2\ln(t(t+1)/\delta)}{|S|}] \geq 1-\delta$$

By definition of $\hat{f}, \mathrm{err}(\hat{f},S)=0$, and therefore

$$Prb[err(\hat{f}) \leq \frac{2\ln(t(t+1)/\delta)}{|S|}] \geq 1-\delta$$