COMP GI13 / M050: Introduction to Statistical Learning Theory 2018/2017

Assignment

Due Date: **5 Jan 2018** *Student Name:* `Yuan Zhang`

Problem 1

1.0

$$(w^*, b^*) = \operatorname*{argmin}_{w \in R^d, b \in R} \left\| y - Xw - b * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_2^2 + n\lambda \|w\|_2^2$$

1.1

In 1.0 we have proved that

$$(w^*, b^*) = \operatorname*{argmin}_{w \in R^d, b \in R} \left\| y - Xw - b * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_2^2 + n\lambda \|w\|_2^2$$

So

$$w^* = \operatorname*{argmin}_{w \in R^d} \left(\min_{b \in R} \left\| y - Xw - b * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_2^2 + n\lambda \|w\|_2^2 \right)$$

Let $L = \left\| y - Xw - b * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_2^2 + n\lambda \|w\|_2^2$

Set the derivative of L wrt b equals to 0

$$y - Xw - b^* * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 0$$

$$\therefore b^* = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \langle w, x_i \rangle \big) = \overline{y} - \overline{X} w$$

$$\begin{aligned}
\therefore w^* &= \underset{w \in R^d}{\operatorname{argmin}} \left\| y - Xw - (\overline{y} - \overline{X}w) * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_2^2 + n\lambda \|w\|_2^2 \\
&= \underset{w \in R^d}{\operatorname{argmin}} \left\| y - \overline{y} * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} - Xw + \overline{X}w * \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_2^2 + n\lambda \|w\|_2^2 \\
&= \underset{w \in R^d}{\operatorname{argmin}} \left\| y_c - X_c w \right\|_2^2 + n\lambda \|w\|_2^2 \\
&= \underset{w \in R^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \big(y_i^c - \langle w, x_i^c \rangle\big)^2 + n\lambda \|w\|_2^2
\end{aligned}$$

So w^* also solves $\min\limits_{w \in R^d} \frac{1}{n} \sum\limits_{i=1}^{n} \big(y_i^c - \langle w, x_i^c \rangle\big)^2 + n\lambda \|w\|_2^2$

1.2

$$\because w^* = \underset{w \in R^d}{\operatorname{argmin}} \left\| y_c - X_c w \right\|_2^2 + n\lambda \|w\|_2^2$$

Let $L = \left\| y_c - X_c w \right\|_2^2 + n\lambda \|w\|_2^2$ and set the derivative of L wrt w equals to 0

$$\therefore w^* = (X_c^T X_c + n\lambda I_n)^{-1} X_c^T y_c$$
$$b^* = \overline{y} - \overline{X} w^*$$

1.3

In 1.2 we get closed form of $w^* = (X_c^T X_c + n\lambda I_n)^{-1} X_c^T y_c$
which means $X_c^T X_c w + n\lambda w - X_c^T y_c = 0$

$$w^* = \frac{X_c^T y_c - X_c^T X_c w}{n\lambda} = X_c^T \frac{(y_c - X_c w)}{n\lambda} = X_c^T * c = \sum_{i=1}^{m} c_i x_i$$

$$\langle w^*, x \rangle = \langle \sum_{i=1}^{m} c_i x_i, x \rangle = \sum_{i=1}^{m} \langle x, x_i \rangle c_i$$

Then we can substitute the inner product $\langle x_i, x \rangle$ with $k(x_i, x)$.

1.4

Let $\gamma = \begin{bmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_n \end{bmatrix}$

$$w^* = \underset{w \in R^d}{\operatorname{argmin}} \left\| \sqrt{\gamma} * (y - Xw) \right\|_2^2 + \lambda \|w\|_2^2$$

1.5

Let $L = \left\| \sqrt{\gamma} * (y - Xw) \right\|_2^2 + \lambda \|w\|_2^2$ and set the derivative of L wrt w equals to 0

$$\begin{aligned} \frac{\partial \big((y - Xw)^T \gamma (y - Xw) + \lambda w^T w \big)}{\partial w} &= 0 \\ -2X^T \gamma y + X^T \gamma X w^* + 2\lambda w^* &= 0 \\ w^* &= (X^T \gamma X + \lambda I_n)^{-1} X^T \gamma y \end{aligned}$$

1.6

From 1.1 we know that w^* also solves centered data without bias. So we substitute X and y with X_c and y_c and get

$$w^* = (X_c^T \gamma X_c + \lambda I_n)^{-1} X_c^T \gamma y_c$$

$$\begin{aligned} b^* &= \sum_{i=1}^{n} \gamma_i (y_i - \langle w^*, x_i \rangle) \\ &= (y - Xw^*)^T \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix} \end{aligned}$$

Problem 2

2.0

$$\|M\|_{op} = \sup_{w \in R^d, \|w\|_2 \le 1} \|Mw\|_2$$

The SVD of M is $M = U \Sigma V^*$, u_i and v_i is the column vector of U and V.

$$w = \sum a_i v_i, \quad (\sum a_i^2 \le 1)$$

$$\begin{aligned}
\therefore \|Mw\|_2 =& \|\sum_{i=1} a_i M v_i\|_2 \\
=& \|\sum_{i=1} a_i \sigma_i u_i\|_2 \\
\leq& \sum_{i=1} \|a_i \sigma_i u_i\|_2 \\
=& \sum_{i=1} |a_i \sigma_i| \\
\leq& |\sigma_{max}|
\end{aligned}$$

$$\therefore \|M\|_{op} = \sigma_{max}$$

2.1

$$h(1) = \nabla F(w'), h(0) = \nabla F(w)$$

$$\therefore \nabla F(w') - \nabla F(w) = h(1) - h(0) = \int_0^1 h'(t)dt$$

$$h'(t) = (w' - w)^T \nabla^2 F(w + t(w' - w))$$

$$\begin{aligned}
\therefore \|\nabla F(w') - \nabla F(w)\|_2 =& \left\| \int_0^1 (w' - w)^T \nabla^2 F(w + t(w' - w))dt \right\|_2 \\
\leq& \|w' - w\|_2 \left\| \int_0^1 \nabla^2 F(w + t(w' - w))dt \right\|_{op}
\end{aligned}$$

According to the definition of L-Lipschitz, $L = \left\| \int_0^1 \nabla^2 F(w + t(w' - w))dt \right\|_{op}$

$$L \leq \int_0^1 \left\| \nabla^2 F(w + t(w' - w)) \right\|_{op} dt = \sup_{w \in R^d} \|\nabla^2 F(w)\|_{op}$$

2.2

a)

$$\nabla^2 F(w) = \tfrac{2}{n} \sum_{i=1}^{n} x_i^T x_i = \tfrac{2}{n} X^T X \qquad (x_i \text{ is a row vector})$$

$$\therefore L \leq \|\frac{2}{n} X^T X\|_{op}$$

b)

$$\nabla^2 F(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2 x_i^T x_i e^{y_i x_i^T w}}{(1+e^{y_i x_i^T w})^2}$$

$$\therefore L \leq \| \frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2 x_i^T x_i e^{y_i x_i^T w}}{(1+e^{y_i x_i^T w})^2}\|_{op} \leq \| \frac{1}{4n}\sum_{i=1}^{n} y_i^2 x_i^T x_i \|_{op}$$

2.3

a)

$$\nabla^2 F(w) = \frac{2}{n}X^T X + 2\lambda I_d$$

$$\therefore L \leq \| \frac{2}{n}X^T X + 2\lambda I_d\|_{op}$$

b)

$$\nabla^2 F(w) = \frac{1}{4n}\sum_{i=1}^{n} y_i^2 x_i^T x_i + 2\lambda I_d$$

$$\therefore L \leq \| \frac{1}{4n}\sum_{i=1}^{n} y_i^2 x_i^T x_i + 2\lambda I_d\|_{op}$$

2.4

In 2.2(a), we know that $L \leq \| \frac{2}{n}\sum_{i=1}^{n} x_i^T x_i\|_{op} \leq \frac{2}{n}\sum_{i=1}^{n}\|x_i^T x_i\|_{op} \leq 2\sup_{x_i}\|x_i^T x_i\|_{op}$. Use the conclusion in 2.0, the upper bound of L equals to $2\sigma_{max}(x_i^T x_i)$.
Set $A = x_i^T x_i, Ax = \sigma_{max} x (x$ is the eigenvector of $\sigma_{max})$. Then $Ax = \sum_{i=1} x_i * a_i$, where x_i is the ith element of vector x, a_i is the ith column of A.

$$|\sigma_{max}|\|x\|_2 = \|Ax\|_2 = \| \sum_{i=1} x_i * a_i\|_2 \leq \|x\|_2 \sqrt{\sum_{i=1}\|a_i\|_2^2}$$

This is because of Cauchy Schwarz inequality.

$$\therefore |\sigma_{max}| \leq \sqrt{\sum_{i=1}\|a_i\|_2^2} = \|x_i\|_2^2 \leq C_1^2$$

$$\therefore L \leq 2C_1^2$$

Similarly, for 2.2(b), $L \leq \frac{1}{4}\sup_{x_i, y_i}\|y_i^2 x_i^T x_i\|_{op}$, the upper bound is $\frac{C_1^2 C_2^2}{4}$.

For 2.3(a) $L \leq \sup_{x_i} \|2x_i^T x_i + 2\lambda\|_{op}$. The eigenvalue of 2.3(a) is increased by λ. So the upper bound of L should add 2λ, which is $2C_1^2 + 2\lambda$.

For 2.3(b) $L \leq \sup_{x_i, y_i} \|\frac{1}{4} y_i^2 x_i^T x_i + 2\lambda\|_{op}$. The eigenvalue of 2.3(b) is increased by λ. So the upper bound of L should add 2λ, which is $\frac{C_1^2 C_2^2}{4} + 2\lambda$.

Problem 3

3.1

We use Mathematical Induction to prove this.
1) For $k = 0, w_0 = X^T c_0 = 0$.
2) Assume we already have $w_k = X^T c_k$, we need to prove $w_{k+1} = X^T c_{k+1}$.
($X = [x_1, \ldots, x_n]^T$, while x_i is a column vector.)
Since $c_{k+1} = c_k - \gamma \frac{2}{n} \bar{\ell}(K^T c_k)$

$$
\begin{aligned}
X^T c_{k+1} &= X^T c_k - \gamma \frac{2}{n} X^T \bar{\ell}(K^T c_k) \\
&= w_k - \gamma \frac{2}{n} X^T \bar{\ell}(X w_k) \\
&= w_k - \gamma \frac{2}{n} \sum_{i=1}^{n} \ell'(\langle w_k, x_i \rangle, y_i) x_i \\
&= w_k - 2\gamma \nabla F(w)
\end{aligned}
$$

$w_{k+1} = w_k - 2\gamma \nabla F(w)$can be seen as the definition of the Gradient Descent Algorithm.

$$\therefore X^T c_{k+1} = w_{k+1}$$

As a result, we have proved that for any $k \in N$, there is c_k so that $w_k = X^T c_k$.

3.2

a)

$$\bar{\ell}(K^T c_k) = 2((\langle w, x_1 \rangle, y_1), \ldots, (\langle w, x_n \rangle, y_n)))^T$$

$$
\begin{aligned}
c_{k+1} &= c_k - \gamma\frac{2}{n}\overline{\ell}(K^T c_k) \\
&= c_k - \gamma\frac{4}{n}(((\langle w, x_1\rangle - y_1), \ldots, (\langle w, x_n\rangle - y_n)))^T \\
&= c_k - \gamma\frac{4}{n}\left((-y_1 + \sum_{i=1}^{n}\langle x_i, x_1\rangle c_{ki}), \ldots, (-y_n + \sum_{i=1}^{n}\langle x_i, x_n\rangle c_{ki})\right)^T
\end{aligned}
$$

We can compute c_k with iterations.
With 2.2(a) we know that Lipschitz constant is $\frac{2}{n}\|K\|_{op}$
From the slides we know that if we let $2\gamma = 1/L$, then

$$
F(w^k) - F(w^*) \leq \frac{L}{2k}\|w^*\|_2
$$

Since L only depends on K, so γ only depends on K.

b)

$$
\begin{aligned}
c_{k+1} &= c_k - \gamma\frac{2}{n}\left(\frac{-y_1}{1+e^{y_1\langle w_k, x_1\rangle}}, \ldots, \frac{-y_n}{1+e^{y_n\langle w_k, x_n\rangle}}\right)^T \\
&= c_k - \gamma\frac{2}{n}\left(\frac{-y_1}{1+e^{y_1\sum_{i=1}^{n}\langle x_i, x_1\rangle c_{ki}}}, \ldots, \frac{-y_n}{1+e^{y_n\sum_{i=1}^{n}\langle x_i, x_n\rangle c_{ki}}}\right)^T
\end{aligned}
$$

We can compute c_k with iterations.
The Lipschitz constant relies on y, so that γ also depends on y apart from K.

Problem 4

4.1

$$
\begin{aligned}
R(c) &= P_{(x,y)\sim\rho}(c(x) \neq y) \\
&= \frac{Area_{(x,y)\sim\rho, c(x)\neq y}(x, y)}{Area_{(x,y)\sim\rho}(x, y)} \\
&= \frac{\int_{X*Y} 1_{c(x)=y}\, d\rho(x, y)}{\int_{X*Y} d\rho(x, y)} \\
&= \int_{X*Y} 1_{c(x)=y}\, d\rho(x, y)
\end{aligned}
$$

4.2

a) $\ell(f(x), y) = (f(x) - y)^2$

$$\begin{aligned}\mathcal{E}(f) &= \int_X \int_Y (f(x) - y)^2 d\rho(y|x) d\rho_X(x) \\ &= \int_X P(1|x)(f(x) - 1)^2 + P(-1|x)(f(x) + 1)^2 d\rho_X(x)\end{aligned}$$

Calculate the derivative wrt f

$$2P(1|x)(f(x) - 1) + 2P(-1|x)(f(x) + 1) = 0$$

$$f(x) = 2P(1|x) - 1$$

b) $\ell(f(x), y) = exp(-yf(x))$

$$\begin{aligned}\mathcal{E}(f) &= \int_X \int_Y exp(-yf(x)) d\rho(y|x) d\rho_X(x) \\ &= \int_X P(1|x)exp(-f(x)) + P(-1|x)exp(f(x)) d\rho_X(x)\end{aligned}$$

Calculate the derivative wrt f

$$P(1|x)exp(-f(x)) + P(-1|x)exp(f(x)) = 0$$

$$f(x) = \frac{1}{2} ln \frac{P(1|x)}{1 - P(1|x)}$$

c) $\ell(f(x), y) = log(1 + exp(-yf(x)))$

$$\begin{aligned}\mathcal{E}(f) &= \int_X \int_Y log(1 + exp(-yf(x))) d\rho(y|x) d\rho_X(x) \\ &= \int_X P(1|x)log(1 + exp(-f(x))) + P(-1|x)log(1 + exp(f(x))) d\rho_X(x)\end{aligned}$$

Calculate the derivative wrt f

$$P(1|x)\frac{-e^{-f(x)}}{1+exp(-f(x))}+P(-1|x)\frac{e^{f(x)}}{1+exp(f(x))}=0$$

$$f(x)=ln\frac{P(1|x)}{1-P(1|x)}$$

d) $\ell(f(x),y)=max(0,1-yf(x))$

$$\begin{aligned}\mathcal{E}(f)&=\int_X\int_Y max(0,1-yf(x))d\rho(y|x)d\rho_X(x)\\&=\int_X P(1|x)max(0,1-f(x))+P(-1|x)max(0,1+f(x))d\rho_X(x)\end{aligned}$$

Since max is convex but not differentiable. If f(x)<-1 or f(x)>1, then truncation of f at -1 or 1 will give a lower loss. So f(x) $\in[-1,1]$.

$$\begin{aligned}\mathcal{E}(f)&=\int_X P(1|x)(1-f(x))+P(-1|x)(1+f(x))d\rho_X(x)\\&=\int_X 1+(1-2P(1|x))f(x))d\rho_X(x)\end{aligned}$$

We can observe that $f(x)=sign(P(1|x)-1/2)$ to give $\epsilon(f)$ minimum value.

4.3

$$\begin{aligned}R(c)&=\int_X\int_Y 1_{c(x)=y}d\rho(y|x)d\rho_X(x)\\&=\int_X P(1|x)1_{c(x)=1}+P(-1|x)1_{c(x)=-1}d\rho_X(x)\end{aligned}$$

Since 0-1 loss function is convex but not differentiable.

If c(x)$\neq\pm1$, the loss $R(c)=\int_X P(1|x)+P(-1|x)d\rho_X(x))$.

Compare to c(x)=1 (the loss $R(c) = \int_X P(-1|x)d\rho_X(x)$)

or c(x)=-1 (the loss $R(c) = \int_X P(1|x)d\rho_X(x)$), they will give a lower loss. So c(x) =1 or -1.

We can observe that $c(x) = sign(P(1|x) - 1/2)$ to give R(c) minimum value.

4.4

In 4.3 we have proved that $c^*(x) = sign(P(1|x) - 1/2)$.

In 4.2(a) we have proved that $f^*(x) = 2p(1|x) - 1$.

Then d(x) = sign(x) will give us the Fisher consistent. Namely, $c^*(x) = d(f^*(x))$.

4.5

4.5.1 $|R(sign(f)) - R(sign(f_*))| = \int_{X_f} |f_*(x)|d\rho_X(x)$

$$\begin{aligned}|R(sign(f)) - R(sign(f_*))| = \bigg| &\int_X P(1|x)1_{sign(f(x))=1} + P(-1|x)1_{sign(f(x))=-1}d\rho_X(x) \\ &- \int_X P(1|x)1_{sign(f_*(x))=1} + P(-1|x)1_{sign(f_*(x))=-1}d\rho_X(x)\bigg|\end{aligned}$$

For $x \in X \setminus X_f$, $sign(f(x)) = sign(f_*(x))$, no contributions to the integral.

$$\begin{aligned}\therefore |R(sign(f)) - R(sign(f_*))| = \bigg| &\int_{X_f} P(1|x)1_{sign(f(x))=1} + P(-1|x)1_{sign(f(x))=-1}d\rho_X(x) \\ &- \int_{X_f} P(1|x)1_{sign(f_*(x))=1} + P(-1|x)1_{sign(f_*(x))=-1}d\rho_X(x)\bigg|\end{aligned}$$

For x that satisfies $sign(f_*(x)) = 1$, $x \in X_f$, then $sign(f(x)) = -1$.

$$\begin{aligned}|R(sign(f)) - R(sign(f_*))|_{sign(f_*(x))=1} &= \left|\int_{X_f, sign(f_*(x))=1} P(1|x) - P(-1|x)d\rho_X(x)\right| \\ &= \left|\int_{X_f, sign(f_*(x))=1} 2P(1|x) - 1d\rho_X(x)\right|\end{aligned}$$

In 4.2(a), we have proved that $f_*(x) = 2P(1|x) - 1$.

$$
\begin{aligned}
\therefore |R(sign(f)) - R(sign(f_*))|_{sign(f_*(x))=1} &= \left| \int_{X_f, sign(f_*(x))=1} f_*(x) d\rho_X(x) \right| \\
&= \int_{X_f, sign(f_*(x))=1} f_*(x) d\rho_X(x) \\
&= \int_{X_f, sign(f_*(x))=1} |f_*(x)| d\rho_X(x)
\end{aligned}
$$

For x that satisfies $sign(f_*(x)) = -1$, $x \in X_f$, then $sign(f(x)) = 1$.

$$
\begin{aligned}
|R(sign(f)) - R(sign(f_*))|_{sign(f_*(x))=-1} &= \left| \int_{X_f, sign(f_*(x))=-1} -P(1|x) + P(-1|x) d\rho_X(x) \right| \\
&= \left| \int_{X_f, sign(f_*(x))=-1} -2P(1|x) + 1 d\rho_X(x) \right|
\end{aligned}
$$

In 4.2(a), we have proved that $f_*(x) = 2P(1|x) - 1$.

$$
\begin{aligned}
\therefore |R(sign(f)) - R(sign(f_*))|_{sign(f_*(x))=-1} &= \left| \int_{X_f, sign(f_*(x))=-1} -f_*(x) d\rho_X(x) \right| \\
&= \int_{X_f, sign(f_*(x))=-1} -f_*(x) d\rho_X(x) \\
&= \int_{X_f, sign(f_*(x))=-1} |f_*(x)| d\rho_X(x)
\end{aligned}
$$

Combine these two parts, $|R(sign(f)) - R(sign(f_*))| = \int_{X_f} |f_*(x)| d\rho_X(x)$.

4.5.2 $\int_{X_f} |f_*(x)| d\rho_X(x) \leq \int_{X_f} \left| f_*(x) - f(x) \right| d\rho_X(x) \leq \sqrt{\mathbb{E}(|f_*(x) - f(x)|^2)}$

For $x \in X_f$, $sign(f_*(x)) \neq sign(f(x))$, $\therefore |f_*(x)| \leq \left| f_*(x) - f(x) \right|$.

$$
\therefore \int_{X_f} |f_*(x)| d\rho_X(x) \leq \int_{X_f} \left| f_*(x) - f(x) \right| d\rho_X(x)
$$

$$
\because X_f \subseteq X \quad \therefore \int_{X_f} \left| f_*(x) - f(x) \right| d\rho_X(x) \leq \int_X \left| f_*(x) - f(x) \right| d\rho_X(x)
$$

According to Cauchy Schwarz inequality,

$$
\int_X \left| f_*(x) - f(x) \right| d\rho_X(x) \leq \sqrt{\int_X |f_*(x) - f(x)|^2 d\rho_X(x)} = \sqrt{\mathbb{E}(|f_*(x) - f(x)|^2)}
$$

4.5.3 $\mathcal{E}(f) - \mathcal{E}(f_*) = \mathbb{E}(|f_*(x) - f(x)|^2)$

$$LEFT = \mathcal{E}(f) - \mathcal{E}(f_*) = \int_X \int_Y (f(x) - y)^2 - (f_*(x) - y)^2 d\rho(y|x) d\rho_X(x) -$$
$$= \int_X P(1|x)(f(x) - 1)^2 + P(-1|x)(f(x) + 1)^2 - P(1|x)(f_*(x) - 1)^2 - P(-1|x)(f_*(x) + 1)^2 d\rho_X(x)$$

$$RIGHT = \mathbb{E}(|f_*(x) - f(x)|^2) = \int_X |f_*(x) - f(x)|^2 d\rho_X(x)$$

So we need to prove the items in the integral equal, which means we need to prove: $P(1|x)(f(x) - 1)^2 + P(-1|x)(f(x) + 1)^2 - P(1|x)(f_*(x) - 1)^2 - P(-1|x)(f_*(x) + 1)^2 = |f_*(x) - f(x)|^2$
Let $P(1|x) = p$, then $P(-1|x) = 1 - p$. In 4.2(a), we have proved that $f_*(x) = 2p - 1$.

$$\begin{aligned} LEFT &= p(f(x) - 1)^2 + (1 - p)(f(x) + 1)^2 - p(2p - 2)^2 - (1 - p)(2p)^2 \\ &= -4pf(x) + 4p^2 - 4p + f(x)^2 + 2f(x) + 1 \\ &= (f(x) + 1 - 2p)^2 \\ &= RIGHT \end{aligned}$$

Problem 5

5.1

Let X_i be a random variable following Bernoulli Distribution. So X_i is bounded in [0,1].

$$\mathbb{P}(X_i) = \begin{cases} p, & X_i = 1 \quad (c(x_i) \neq y_i) \\ 1 - p, & X_i = 0 \quad (c(x_i) = y_i) \end{cases}$$

Use Hoeffding's inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_i) \geq \epsilon\right) \leq exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right)$$

$$\because a_i = 0, b_i = 1 \therefore \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_i) \geq \epsilon\right) \leq exp(-2n\epsilon^2)$$

$$\Longrightarrow \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq n(\epsilon + p)\right) \leq exp(-2n\epsilon^2)$$

Let $m = n(p + \epsilon)$, then $\epsilon = (m - pn)/n$.

$$\Longrightarrow \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq m\right) \leq exp(-2(m - pn)^2/n)$$

$$\implies 1 - \mathbb{P}\Bigg(\sum_{i=1}^{n} X_i < m\Bigg) \leq exp(-2(m - pn)^2/n)$$

$$\implies \mathbb{P}\Bigg(\sum_{i=1}^{n} X_i < m\Bigg) \geq 1 - exp(-2(m - pn)^2/n)$$

And we have $\mathbb{P}\Bigg(\sum_{i=1}^{n} X_i = m\Bigg) = \left(\begin{array}{c} m \\ n \end{array}\right) p^m (1-p)^{n-m}$

$$\implies \mathbb{P}\Bigg(\sum_{i=1}^{n} X_i \leq m\Bigg) \geq 1 - exp(-2(m - pn)^2/n) + \left(\begin{array}{c} m \\ n \end{array}\right) p^m (1-p)^{n-m}$$

So the lower bound of for the probability that c makes at most m mistakes on S is

$$1 - exp(-2(m - pn)^2/n) + \left(\begin{array}{c} m \\ n \end{array}\right) p^m (1-p)^{n-m}$$