10702/36702 Statistical Machine Learning, Spring 2008
Homework 2 Solutions

February 20, 2008

1 [15 points]

(a) Let π_n be a sequence of priors and $\tilde{\theta}_n$ the corresponding Bayes estimators. Suppose that

$$\int R(\theta, \tilde{\theta}_n)\pi_n(\theta)d\theta \to c$$

for some finite c. Suppose that $\hat{\theta}$ is an estimator such that

$$\sup_\theta R(\theta, \hat{\theta}) \leq c$$

Show that $\hat{\theta}$ is minimax.

★ **SOLUTION:** For any estimator T

$$\begin{aligned}\sup_\theta R(\theta, T) &\geq \int R(\theta, T)\pi_n(\theta)d\theta \\ &\geq \int R(\theta, \tilde{\theta}_n\pi_n(\theta)d\theta\end{aligned}$$

Let n goes to infinity on both sides of the inequality, we have

$$\lim_{n\to\infty} \sup_\theta R(\theta, T) = \sup_\theta R(\theta, T) \geq \lim_{n\to\infty} \int R(\theta, \tilde{\theta}_n)\pi_n(\theta)d\theta = c \geq \sup_\theta R(\theta, \hat{\theta})$$

Therefore, $\hat{\theta}$ is minimax.

(a) Let $X \sim N(\theta, 1)$. Show that $\hat{\theta} = X$ is minimax.
Hint: Let π_n be $N(0, n)$. Check that

$$\int R(\theta, \hat{\theta}_n)\pi_n(\theta)d\theta \to 1$$

Next show that $R(\theta, X) = 1$. Consider from part (a) that X is minimax.

★ **SOLUTION:** Let π_n be $N(0, n)$. The posterior distribution

$$\begin{aligned}\pi(\theta|X = x) &\propto \exp(-\frac{(x-\theta)^2}{2}) \cdot \exp(-\frac{\theta^2}{2n}) \\ &\propto \exp(-\frac{(n+1)(\theta - \frac{n}{n+1}x)^2}{2n}\end{aligned}$$

Therefore, $\pi(\theta|X=x) \sim N(\frac{n}{n+1}X, \frac{n}{n+1})$, and the Bayes estimator $\tilde{\theta}_n = \frac{n}{n+1}X$. The risk

$$R(\theta, \tilde{\theta}_n) = E_\theta(\theta - \tilde{\theta}_n)^2 = (E_\theta\tilde{\theta}_n - \theta)^2 + Var_\theta(\tilde{\theta}_n) = \frac{\theta^2 + n^2}{(n+1)^2}$$

The Bayes risk

$$\int R(\theta, \tilde{\theta}_n)\pi_n(\theta)d\theta = \int \frac{\theta^2 + n^2}{(n+1)^2} \cdot \frac{\exp(-\frac{\theta^2}{2n})}{\sqrt{2\pi n}} d\theta = \frac{n + n^2}{(n+1)^2} = \frac{n}{n+1}$$

Therefore, as $n \to \infty$, the Bayes risk goes to 1. On the other hand,

$$R(\theta, X) = (E_\theta(X) - \theta)^2 + Var_\theta X = 1$$

Therefore, $\sup_\theta R(\theta, X) = 1$. According to (a), X is minimax.

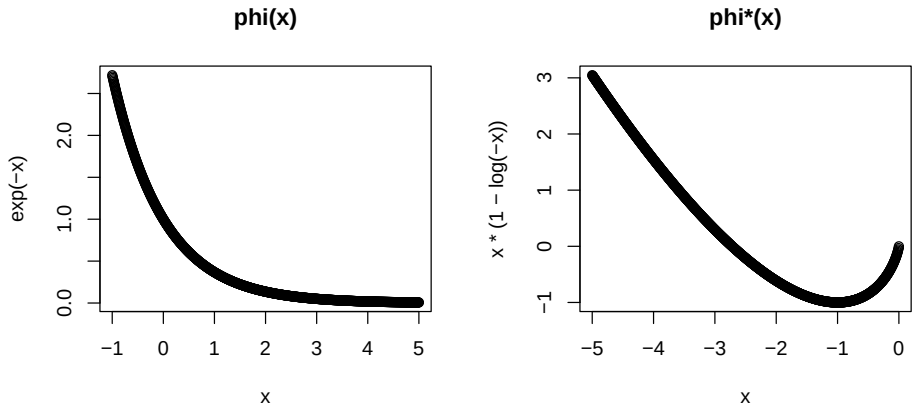
2 [16 points]

The following is a list of some loss functions commonly used for large-margin classification algorithms. For each loss function $\phi(x)$ determine whether ϕ is a convex function, and then calculate its conjugate $\phi^\star$. Plot ϕ and $\phi^\star$.

(a) Exponential loss: $\phi(x) = exp(-x)$

★ **SOLUTION:** $\phi''(x) = exp(-x) \succeq 0 \quad \forall x$. Hence $\phi(x)$ is convex.
Conjugate function: $\phi^\star(y) = \sup_x(xy - exp(-x))$.
For $y > 0$, $\phi^\star(y)$ is unbounded.
For $y = 0$, $\phi^\star(y) = 0$.
For $y < 0$, $\frac{\partial}{\partial x}(xy - exp(-x)) = y - exp(-x) = 0$.
Or, $x^\star = -log(-y)$. Substituting x, we get $\phi^\star(y) = -ylog(-y) + y$.

$$\phi^\star(y) = \begin{cases} \infty & y > 0 \\ 0 & y = 0 \\ -ylog(-y) + y & y < 0 \end{cases}$$



(b) Truncated quadratic loss: $\phi(x) = [\max(1 - x, 0)]^2$

★ **SOLUTION:** $1-x$ and 0 are both convex. Hence $\max[(1-x),0]$ is also convex. Square of convex function also convex. Hence $\phi(x)$ convex.

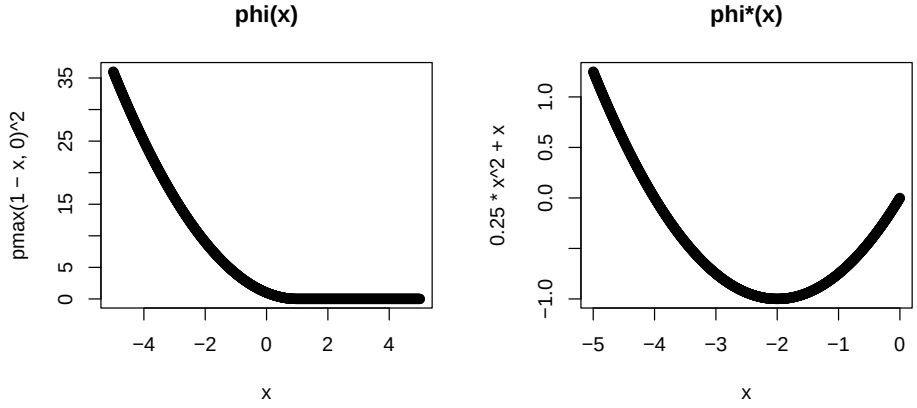
Conjugate computation:

For $y > 0$, $\phi^\star(y)$ is unbounded since $x^\star \to \infty$.

Consider $y \leq 0$. Then for $x \leq 1$ we have $\phi(x) = (1-x)^2$ and $\phi^\star(y) = \sup(xy - (1-x)^2)$. To find the maximum we differentiate to get $y + 2(1-x) = 0$ or $x^\star = 1 + y/2$. Substituting for x in $\phi^\star(y)$ we get $\phi^\star(y) = (1+y/2)y - y^2/4 = y^2/4 + y$. For $x > 1$, the conjugate is unbounded.

Hence we have the conjugate as:

$$\phi^\star(y) = \begin{cases} y^2/4 + y & y \leq 0 \\ \infty, & otherwise \end{cases}$$



(c) Hinge loss: $\phi(x) = \max(1-x, 0)$

★ **SOLUTION:** $\phi(x)$ is convex. see part (b).

Conjugate computation:

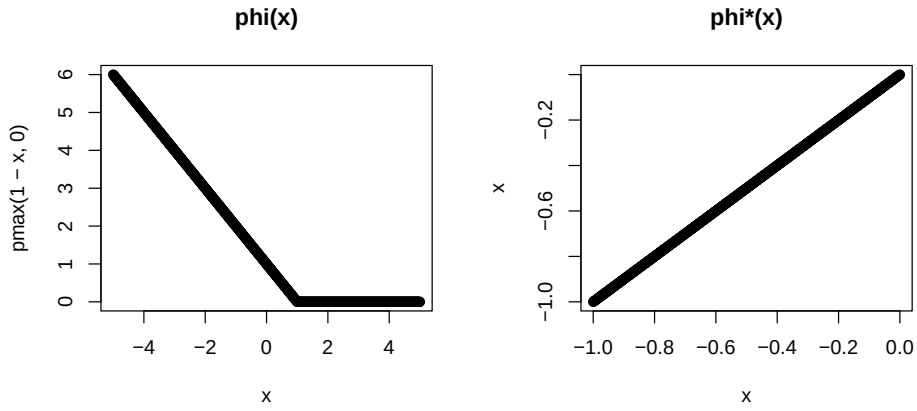
For $y > 0$, $\phi^\star(y)$ is unbounded since $x^\star \to \infty$.

For $y < -1$, $\phi^\star(y)$ is unbounded since $x^\star \to -\infty$.

Consider the case when $-1 \leq y \leq 0$. We have $\phi(x) = (1-x)$ when $x \leq 1$ and 0 otherwise. Hence, $\phi^\star(y) = \sup(xy - (1-x))$. Differentiating we get, $y + 1 = 0$ or $y = -1$. Or, $\phi^\star(y) = y$

Hence we have the conjugate as:

$$\phi^\star(y) = \begin{cases} y & -1 \leq y \leq 0 \\ \infty, & otherwise \end{cases}$$



(d) Sigmoid loss: $\phi(x) = 1 - \tanh(\kappa x)$, for fixed $\kappa > 0$

★ **SOLUTION:** $\phi''(x) = 2\kappa^2 \text{sech}^2(\kappa x)\tanh(\kappa x)$. Hence, $\phi''(x)$ has the same sign as $\tanh(\kappa x)$ which can be positive or negative. Hence $\phi(x)$ is not convex.

Conjugate computation:

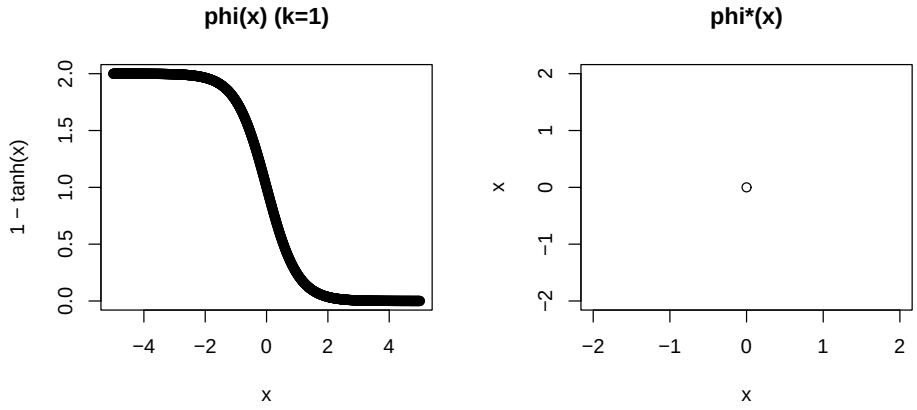
For $y > 0$, $\phi^\star(y)$ is unbounded since $x^\star \to \infty$.

For $y < 0$, $\phi^\star(y)$ is unbounded since $x^\star \to -\infty$.

Only when $y = 0$, we have $\phi^\star(y) = -1 + \tanh(\kappa x)$. Since hyperbolic tan is bound between -1 and 1, we have $\phi^\star(0) = 0$.

Hence we have the conjugate as:

$$\phi^\star(y) = \begin{cases} 0 & y = 0 \\ \infty, & otherwise \end{cases}$$



3 [14 points]

If $f(x, y) = f_1(x) + f_2(y)$, with f_1 and f_2 convex, show that

$$f^\star(x, y) = f_1^\star(x) + f_2^\star(y)$$

Does this hold if f_1 and f_2 are not convex?

★ **SOLUTION:**

$$\begin{aligned} f^\star(x,y) &= \sup(< u,v >< x,y >' - f(u,v)) \\ &= \sup(ux + vy - f(u) - f(v)) \\ &= \sup(ux - f(u)) + \sup(vy - f(y)) \\ &= f^\star(x) + f^\star(y) \end{aligned}$$

Since we didn't need the convexity condition, this also holds for non-convex functions.

4 [15 points]

The following is called the *probit regression model*. Suppose $Y \in \{0,1\}$ is a random variable given by

$$Y = \begin{cases} 1 & a^T X + b + V \leq 0 \\ 0 & a^T X + b + V > 0 \end{cases}$$

where $X \in \mathbb{R}^p$ is a vector of explanatory variables and $V \sim N(0,1)$ is a latent (unobserved) random variable. Formulate the maximum likelihood estimation problem of estimating a and b, given data consisting of pairs (X_i, Y_i), $i = 1, \ldots, n$, as a convex optimization problem.

★ **SOLUTION:**

$$P(Y = 1|X) = P(a^T X + b + V \leq 0) = P(V \leq -a^T X - b) = \Phi(-a^T X - b)$$

where $\Phi(\cdot)$ is the standard normal cdf. Therefore, $P(Y = 0|X) = 1 - P(Y = 1|X) = \Phi(a^T X + b)$.

The log-likelihood

$$\begin{aligned} l(a,b) &= \sum_{i=1}^{n} [Y_i \log(P(Y = 1|X_i)) + (1 - Y_i) \log(P(Y = 0|X_i))] \\ &= \sum_{i=1}^{n} [Y_i \log \Phi(-a^T X - b) + (1 - Y_i) \log \Phi(a^T X + b)] \end{aligned} \tag{1}$$

According to the notes on log-concavity, $\Phi(\cdot)$ is log-concave. Thus, $\log \Phi(\cdot)$ is a non-decreasing concave function. Since $-a^T X - b$ and $a^T X + b$ are concave functions of a and b, $\log \Phi(-a^T X - b)$ and $\log \Phi(a^T X + b)$ are both concave functions of a and b. According to equation (1), $l(a,b)$ is a non-negative weighted combination of concave functions. Therefore, $l(a,b)$ is a concave function.

The maximum likelihood estimation problem is to maximize the concave function $l(a,b)$, which is equivalent to minimize the convex function $-l(a,b)$, so it is a convex optimization problem.

5 [15 points]

For $x \in \mathbb{R}^n$ define the L_p norm

$$\|x\|_p = (\sum_{j=1}^{n} |x_j|^p)^{1/p}$$

for $p > 0$. Let

$$C = \{x : \|x\|_p \leq 1\}$$

Show that C is convex if and only if $p \geq 1$.

★ **SOLUTION:** First, we prove that if $p \geq 1$, then C is convex. Let $f(y) = y^p$, where $y > 0$ and $p \geq 1$. It is easily verified that $f(y)$ is a non-decreasing convex function of y. Let $g(y) = |y|$, where $y \in \mathbb{R}$. $g(y)$ is also a convex function. Therefore, $f(g(y)) = |y|^p$ is a convex function, where $y \in \mathbb{R}$ and $p \geq 1$.

$\forall x^1, x^2 \in C$, $\sum_{j=1}^n |x_j^1|^p \leq 1$ and $\sum_{j=1}^n |x_j^2|^p \leq 1$. According to Jensen's inequality, $\forall c \in [0,1]$

$$\sum_{j=1}^n |cx_j^1 + (1-c)x_j^2|^p \leq \sum_{j=1}^n [c|x_j^1|^p + (1-c)|x_j^2|^p] \leq c\sum_{j=1}^n |x_j^1|^p + (1-c)\sum_{j=1}^n |x_j^2|^p = 1$$

Therefore, $cx^1 + (1-c)x^2 \in C$, and C is convex.

Next, we prove that if $p < 1$, then C is not convex. If $n = 1$, it can be easily verified that C is not convex. If $n > 1$, let $x^1 = [1, 0, \ldots, 0]^T$, $x^2 = [0, 1, \ldots, 0]^T$ and $c = 0.5$.

$$\sum_{j=1}^n |cx_j^1 + (1-c)x_j^2|^p = 2 \times 0.5^p > 2 \times 0.5 = 1$$

Therefore, $cx^1 + (1-c)x^2 \notin C$, and C is not convex.

To summarize, C is convex if and only if $p \geq 1$.

6 [14 points]

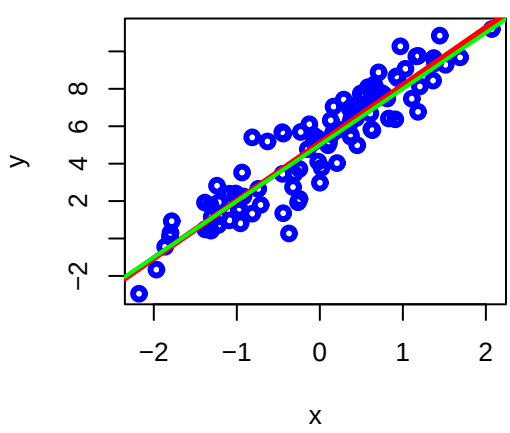
Linear regression in R. Add brief comments to this code, and to the output, to explain what the code does and what the output means.

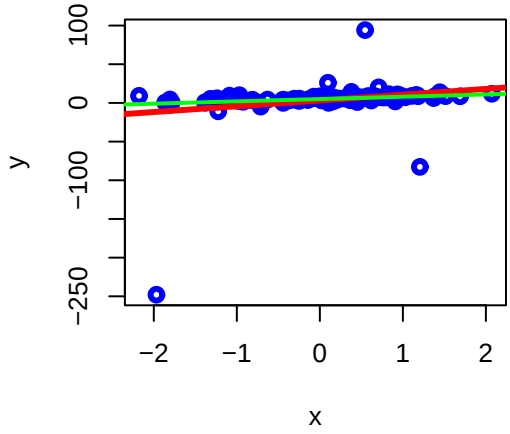
★ **SOLUTION:**

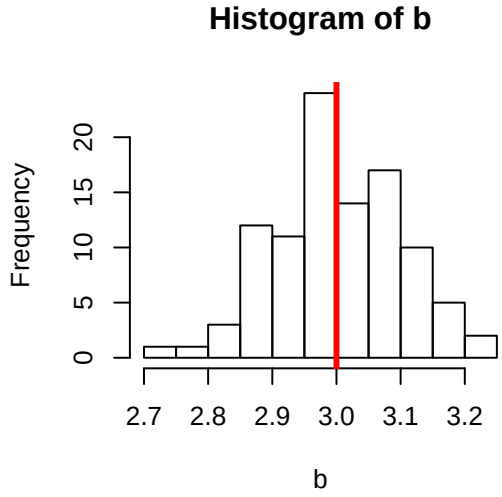
```
par(mfrow=c(2,2),bg='cornsilk')      # plots will be drawn on a 2x2 grid on cornsilk background
n = 100
sigma = 1
x = rnorm(n)                         # generate 'n' random numbers from N(0,1)
x = sort(x)
y = 5 + 3*x + rnorm(n,0,sigma)       # y is linear function of x with normal noise
plot(x,y,col="blue",lwd=3)           # plot (x,y) with blue color
out = lm(y~x)                        # fit a linear model
summary(out)                         # show summary of fitted linear model
abline(out,col="red",lwd=3)          # show predicted values of 'y' with 'red' lines
abline(a=5,b=3,col="green",lwd=2)    # show true value of 'y' with 'green' lines

y = 5 + 3*x + rcauchy(n,0,sigma)     # y is linear function of x with cauchy noise
plot(x,y,col="blue",lwd=3)
out = lm(y~x)
summary(out)
abline(out,col="red",lwd=3)
abline(a=5,b=3,col="green",lwd=2)

nsim = 100
b = rep(0,nsim)
for(i in 1:nsim){                    # generate random x and y 100 times with normal noise
  x = rnorm(n)                       # and record the predicted slope of regression line
  x = sort(x)                        # for each simulation
  y = 5 + 3*x + rnorm(n,0,sigma)
  out = lm(y~x)
```

```
  b[i] = out\$coef[2]
}
summary(b)
hist(b)                              # plot the histogram of predicted slope values
abline(v=3,lwd=3,col="red")          # show true value in the histogram
print(mean((b-3)^2))                 # print MSE value

#####################################################################
# Output of summary(out) with normal noise
#####################################################################
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.91252 -0.63554 -0.04475  0.62950  2.61872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1733     0.1052   49.19   <2e-16 ***
x             3.0534     0.1128   27.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.037 on 98 degrees of freedom
Multiple R-Squared: 0.8821,     Adjusted R-squared: 0.8809
F-statistic: 733.2 on 1 and 98 DF,  p-value: < 2.2e-16

As seen above, the estimates of regression coefficients are 5.17 and
3.05. These are very close to the true values and very significant (<2e-16).
The plot shows that the ranlge of 'y' is very large as compared to normal.
#####################################################################

#####################################################################
# Output of summary(out) with cauchy noise
#####################################################################
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-20.2287  -0.9422   0.1061   1.0327   7.8915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0879     0.3379  15.058  < 2e-16 ***
x             2.6615     0.3683   7.227 1.09e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.356 on 98 degrees of freedom
Multiple R-Squared: 0.3477,     Adjusted R-squared: 0.341
F-statistic: 52.23 on 1 and 98 DF,  p-value: 1.087e-10
```

As seen above, the estimates with cauchy noise are not as good as from normal noise. This is due to the fact that cauchy has heavier tails.







7 [14 points]

Prove the leave-one-out cross-validation identity:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_{(-i)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}}\right)^2$$

★ **SOLUTION:** Consider Z such that

$$Z_j = \begin{cases} Y_j & j \neq i \\ \hat{Y}_j^{(-i)} & j = i \end{cases}$$

SSE $= \sum(Y_j^{(-i)} - Z_j)^2$. This is also the minimal SSE for the regression on Y without the i^{th} data point.

$$\begin{aligned} \text{Hence,} \hat{Y}_i^{(-i)} &= (HZ)_i \\ &= \sum_{k \neq i} H_{ik}Z_k + H_{ii}Z_i \\ &= \sum H_{ik}Y_k - H_{ii}Y_i + H_{ii}\hat{Y}_i^{(-i)} \\ &= (HY)_i - H_{ii}Y_i + H_{ii}\hat{Y}_i^{(-i)} \\ &= \hat{Y}_i - H_{ii}Y_i + H_{ii}\hat{Y}_i^{(-i)} \\ &= \frac{\hat{Y}_i - H_{ii}Y_i}{1 - H_{ii}} \end{aligned}$$

Using this we get,

$$\frac{1}{n}\sum(Y_i - \hat{Y}_i^{(-i)})^2 = \frac{1}{n}\sum\left(\frac{Y_i - Y_iH_{ii} - \hat{Y}_i + H_{ii}Y_i}{1 - H_{ii}}\right)^2 = \frac{1}{n}\sum\left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}}\right)^2$$

Hence proved.