6.867 Exam 1
Fall 2011

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page. Show your work neatly.

Name and MIT ID: _____

Question	Points	Score
1	35	
2	25	
3	25	
4	15	
Total:	100	

1 Estimation and decision theory (35 points)

In all parts of the following question, feel free to leave any complex numerical expressions unevaluated: you can just write them down, and then give them a name if you need to use them later.

You have just bought a copy machine at a garage sale. You know it is one of two possible models, m_1 or m_2, but the tag has fallen off, so you're not sure which.

You do know that m_1 machines have a 0.1 "error" (bad copy) rate and m_2 machines have a 0.2 error rate.

1. (a) (5 points) You use your machine to make 1000 copies, and 140 of them are bad. What is the maximum likelihood estimate of the machine's error rate? Explain why. (Remember that you're sure it's one of those two types of machines).

Solution:
We first solve the MLE of the type of the machine, which we denote by $b \in \{1, 2\}$. Using a particular machine, the number of bad copies, denoted by k, is a random variable, as $k \sim \text{Binomial}(n, p_b)$. Thus,

$$P(k \mid b) = \binom{n}{k} p_b^k (1-p_b)^{n-k} \;\Rightarrow\; \log P(k|b) = \log C + k \log p_b + (n-k)\log(1-p_b).$$

Here, C is the value of n *choose* k. With $n = 1000$, $k = 140$, $p_1 = 0.1$ and $p_2 = 0.2$, we have

$$\log P(k|b=1) = \log C + 140\log(0.1) + 860\log(0.9) = \log C - 412.97$$
$$\log P(k|b=2) = \log C + 140\log(0.2) + 860\log(0.8) = \log C - 417.22$$

We can see that $\log P(k|b=1) > \log P(k|b=2)$, which implies that the MLE of the type of the machine is $\hat{b} = 1$. It follows that the machine's error rate is $p_{\hat{b}} = 0.1$.

(b) (10 points) Looking more closely, you can see part of the label, and so you think that, just based on the label it has a probability 0.2 of being an m_1 type machine and a probability 0.8 of being an m_2 type machine. If you take that to be your prior, and incorporate the data from part a, what is your posterior distribution on the type of the machine?

Solution: Under the condition that the total number of copies that we made is $n = 1000$, the posterior distribution of the type of the machine, denoted by b, is

$$p(b=1 \mid k) = \frac{p(k \mid b=1)p(b=1)}{p(k \mid b=1)p(b=1) + p(k \mid b=2)p(b=2)} = \frac{0.2}{0.2 + 0.8\frac{p(k|b=2)}{p(k|b=1)}}.$$

We note that $\log p(k \mid b=2) - \log p(k \mid b=1) = -4.25$. Hence

$$\frac{p(k \mid b=2)}{p(k|b=1)} = \exp(-4.25) = 0.0142.$$

As a result, we have

$$p(b=1 \mid k) = 0.946, \quad \text{and} \quad p(b=2 \mid k) = 0.054.$$

(c) (5 points) Given that posterior, what is the probability that the next copy will be a failure?

Solution: Given the posterior, the predictive probability of the next copy being bad is

$$p(b = 1 \mid k)p_1 + p(b = 2 \mid k)p_2 = 0.946 \cdot 0.1 + 0.054 \cdot 0.2 = 0.1054.$$

(d) (10 points) You intend to sell this machine on the web. Because it's used, you have to sell it with a warrantee. You can offer a gold or a silver warrantee. If it has a gold warrantee and the buyer runs it for 1000 copies and gets more than 150 bad copies, then you are obliged to pay $1000 in damages; if it has a silver warrantee, you have to pay damages if it generates more than 300 bad copies in 1000 copies. Your maximum reasonable asking price for a machine with a gold warrantee is $300; for a machine with a silver warrantee, it is $100. You can assume the machine will sell at these prices. What type of warrantee should you offer on this machine?

Solution: Let $k = 140$ denote the number of bad copies that we have observed, and k' denote the number of bad copies the machine will generate when the buyer runs it for 1000 new copies. The probability that $k' > 150$ is

$$p(k' > 150 \mid k) \;=\; p(k' > 150 \mid b = 1)p(b = 1 \mid k) + p(k' > 150 \mid b = 2)p(b = 2 \mid k).$$

When $n = 1000$, the binomial distribution is extremely peaky, with most probability mass falling around np. Hence, $p(k' > 150 \mid b = 1) \simeq 0$, and $p(k' > 150 \mid b = 2) \simeq 1$. Hence $p(k' > 150 \mid k) \simeq p(b = 2 \mid k) = 0.054$.

Similarly, we have

$$p(k' > 300 \mid k) \;=\; p(k' > 300 \mid b = 1)p(b = 1 \mid k)+p(k' > 300 \mid b = 2)p(b = 2 \mid k) \simeq 0.$$

Actually, using either machine, it is very unlikely to generate over 300 bad copies for 1000 runs.

Hence, the expected profit of offering gold warrantee is

$$300 - 1000 \cdot p(k' > 150 \mid k) \simeq 300 - 1000 \cdot 0.054 = 246.$$

The expected profit of offering silver warrantee is

$$100 - 1000 \cdot p(k' > 300 \mid k) \simeq 100.$$

Therefore, offering gold warrantee would generate higher expected profit, which is what we should do.

(e) (5 points) Under what conditions would it be better to just throw the machine away, rather than try to sell it?

Solution: We should just throw it away when *the expected profit is zero or even negative* for both warrantee that we can offer. (You can get full points if you see the above, but it is great if you see the following.)

For this particular problem, even for the worst case scenario where we are sure with probability 1 that the machine is the worse one (with error rate 0.2), it is still very unlikely that it produces over 300 bad copies for 1000 runs (you can verify this by computing the CDF). In this (worst) case, it is still profitable to sell the machine with silver warrantee.

2 Weighted least squares regression (25 points)

You are trying to build a classifier with data that you gathered on two different days with two different instruments. We know that data set 1, consisting of n pairs, (x^i, y^i) has a conditional Gaussian distribution

$$y \sim \text{Normal}(x \cdot \theta, \sigma_1^2) \ ,$$

and data set 2, consisting of m pairs (u^i, v^i) has a conditional Gaussian distribution that differs only in the variance:

$$v \sim \text{Normal}(u \cdot \theta, \sigma_2^2) \ ,$$

The parameter vector θ and all of the x^i and u^i are vectors in R^d, and the y^i and v^i are in R.

2. (a) (20 points) Derive the maximum-likelihood estimator for $\theta \in R^d$. You can assume that there is no special θ_0. **We strongly recommend that you do this in matrix-vector form.**

Solution: For dataset 1

$$\log L_1(X;\theta) = \frac{1}{2\sigma_1^2}\sum_{i=1}^{n}(<\theta, x_i> - y_i)^2 + C_1 = \frac{1}{2\sigma_1^2}\|X\theta - Y\|^2 + C_1$$

For dataset 2

$$\log L_1(U;\theta) = \frac{1}{2\sigma_2^2}\sum_{i=1}^{m}(<\theta, u_i> - v_i)^2 + C_2 = \frac{1}{2\sigma_2^2}\|U\theta - V\|^2 + C_2$$

where $X = [x_1, \cdots, x_n]^T, Y = [y_1, \cdots, y_n]^T, V = [v_1, \cdots, v_m]^T, U = [u_1, \cdots, u_m]^T$ and C_1, C_2 are constants from the Gaussian PDF.
The joint objective of MLE is

$$\begin{aligned} J &= \log L_1 + \log L_2 \\ &= \frac{1}{2\sigma_1^2}(X\theta - Y)^T(X\theta - Y) + \frac{1}{2\sigma_2^2}(U\theta - V)^T(U\theta - V) + C_1 + C_2 \\ &= \frac{1}{2}\theta^T\left[\frac{1}{\sigma_1^2}X^TX + \frac{1}{\sigma_2^2}U^TU\right]\theta - \left[\frac{1}{\sigma_1^2}\theta^TX^Ty + \frac{1}{\sigma_2^2}\theta^TU^TV\right] + C \end{aligned}$$

Maximizing J with respect to θ

$$\frac{\partial J}{\partial \theta} = \left[\frac{1}{\sigma_1^2}X^TX + \frac{1}{\sigma_2^2}U^TU\right]\theta - \left[\frac{1}{\sigma_1^2}\theta^TX^Ty + \frac{1}{\sigma_2^2}\theta^TU^TV\right] = 0$$

The solution is

$$\hat{\theta} = \left[\frac{1}{\sigma_1^2}X^TX + \frac{1}{\sigma_2^2}U^TU\right]^{-1}\left[\frac{1}{\sigma_1^2}\theta^TX^Ty + \frac{1}{\sigma_2^2}\theta^TU^TV\right]$$

(b) (5 points) Argue that it makes sense for extreme relative values of σ_1 and σ_2.

Solution: Since σ_1 and σ_2 can be viewed as weighting coefficients, if

$$\begin{cases} \sigma_1 = \sigma_2 & \longrightarrow \text{ unweighted (or equally weighted) least square regression} \\ \sigma_1 \gg \sigma_2 & \longrightarrow \text{ least square regression on } (u^i, v^i) \\ \sigma_1 \ll \sigma_2 & \longrightarrow \text{ least square regression on } (x^i, y^i) \end{cases}$$

3 SVMs with multiple data sources (25 points)

You are still trying to build a classifier with data that you gathered on two different days with two different instruments. You trust the labels of the data gathered with instrument 1 twice as much as the labels of the data gathered with instrument 2. You have lots of friends with different opinions about how to handle this.

We will use $(x^i, y^i), i = 1 \ldots n$ to denote data from instrument 1 (more accurate) and (u^i, v^i) to denote data from instrument 2. Slack variables for the instrument 1 data will be ξ and for the instrument 2 data will be ζ. Lagrange multipliers for the instrument 1 data will be α and for the instrument 2 data will be β.

- Pat suggests that you can insert a multiplier of 2 into the slack penalties for the data points gathered with instrument 1, so that the optimization problem is

$$\min_{\theta,\xi,\zeta} \frac{1}{2}||\theta||^2 + 2c\sum_{i=1}^{n}\xi_i + c\sum_{j=1}^{m}\zeta_j$$

  subject to

$$\begin{aligned} y^i(\theta \cdot x^i + \theta_0) &\geq 1 - \xi_i \quad \text{for all } i \in \{1, \ldots, n\} \\ v^j(\theta \cdot u^j + \theta_0) &\geq 1 - \zeta_j \quad \text{for all } j \in \{1, \ldots, m\} \\ \xi_i &\geq 0 \quad \text{for all } i \in \{1, \ldots, n\} \\ \zeta_j &\geq 0 \quad \text{for all } j \in \{1, \ldots, m\} \end{aligned}$$

- Dana suggests that you can insert a multiplier of 2 into the Lagrange multipliers of data points gathered with instrument 1, so that the dual optimization problem is:

$$\max_{\alpha,\beta} \sum_{i=1}^{n} 2\alpha_i + \sum_{j=1}^{m}\beta_j - 2\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j (x^i \cdot x^j) - 2\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j y^i v^j (x^i \cdot u^j) - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\beta_i\beta_j v^i v^j (u^i \cdot u^j)$$

  subject to

$$\begin{aligned} c \geq 2\alpha_i &\geq 0 \quad \text{for all } i \in \{1, \ldots, n\} \\ c \geq \beta_j &\geq 0 \quad \text{for all } j \in \{1, \ldots, m\} \\ \sum_{i=1}^{n} 2\alpha_i y^i + \sum_{j=1}^{m}\beta_j v^j &= 0 \end{aligned}$$

- Robin suggests that you can duplicate the points that you gathered with instrument 1 in the data set, and then proceed as usual.

3. (25 points) Are these approaches equivalent, in the sense of resulting in the same separator? For each pair, show that they are equivalent or not.

Solution: We begin with deriving the dual form of Pat's suggestion. Using Lagrange multipliers, the new objective function is

$$L(\theta, \theta_0, a, b, e, f) = \frac{1}{2}\|\theta\|^2 + 2c\sum_{i=1}^{n}\xi_i + c\sum_{j=1}^{m}\zeta_j - \sum_{i=1}^{n} a_i\left[y^i(\theta\cdot x^i + \theta_0) - 1 + \xi_i\right] - \sum_{i=1}^{m} b_i\left[v^i(\theta\cdot u^i + \theta_0) - 1 + \zeta_i\right] - \sum_{i=1}^{n} e_i\xi_i - \sum_{i=1}^{m} f_i\zeta_i$$

subject to

$$\begin{aligned}
&a_i \geq 0, b_i \geq 0, e_i \geq 0, f_i \geq 0\\
&y^i(\theta\cdot x^i + \theta_0) - 1 + \xi_i \geq 0\\
&v^i(\theta\cdot u^i + \theta_0) - 1 + \zeta_i \geq 0\\
&e_i\xi_i = 0, f_i\zeta_i = 0\\
&a_i\left[y^i(\theta\cdot x^i + \theta_0) - 1 + \xi_i\right] = 0\\
&b_i\left[v^i(\theta\cdot u^i + \theta_0) - 1 + \zeta_i\right] = 0
\end{aligned}$$

Optimizing $\theta, \theta_0, \xi_i, \zeta_i$

$$\frac{\partial L}{\partial \theta} = 0 \implies \theta = \sum_{i=1}^{n} a_i y^i x^i + \sum_{i=1}^{m} b_i v^i u^i \tag{1}$$

$$\frac{\partial L}{\partial \theta_0} = 0 \implies \sum_{i=1}^{n} a_i y^i + \sum_{i=1}^{m} b_i v^i = 0 \tag{2}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \implies a_i = 2c - e_i \tag{3}$$

$$\frac{\partial L}{\partial \zeta_i} = 0 \implies b_i = c - f_i \tag{4}$$

Since $a_i \geq 0, b_i \geq 0, e_i \geq 0, f_i \geq 0$, Eq. 3 and 4 become

$$\begin{aligned}
&2c \geq a_i \geq 0\\
&c \geq b_i \geq 0
\end{aligned}$$

Use these results to eliminate $\theta, \theta_0, \xi_i, \zeta_i$ from the Lagrangian, we obtain the dual Lagrangian in the form

$$\bar{L}(a, b) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{m} b_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y^i y^j (x^i\cdot x^j) - \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j y^i v^j (x^i\cdot u^j) - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} b_i b_j v^i v^j (u^i\cdot u^j)$$

with constraints

$$
\begin{aligned}
&2c \geq a_i \geq 0 \\
&c \geq b_i \geq 0 \\
&\sum_{i=1}^{n} a_i y^i + \sum_{i=1}^{m} b_i v^i = 0
\end{aligned}
$$

Let $\alpha_i = \frac{a_i}{4}$, $\beta_i = b_i$, the dual form can be rewritten as

$$
\begin{aligned}
\max_{\alpha,\beta} \sum_{i=1}^{n} 4\alpha_i + \sum_{i=1}^{m} \beta_i - 8 \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j) - 4 \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j y^i v^j (x^i \cdot u^j) \\
- \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \beta_i \beta_j v^i v^j (u^i \cdot u^j)
\end{aligned}
$$

subject to

$$
\begin{aligned}
&c \geq 2\alpha_i \geq 0 \\
&c \geq \beta_i \geq 0 \\
&\sum_{i=1}^{n} 4\alpha_i y^i + \sum_{i=1}^{m} \beta_i v^i = 0
\end{aligned}
$$

It is clear that Dana's suggestion is different from Pat's. Therefore, Dana $\neq$ Pat.

If we use the same slack variable for points after duplicating, Robin's suggestion can be written as

$$
\min_{\theta,\xi,\zeta} \frac{1}{2}||\theta||^2 + c \sum_{i=1}^{n} \xi_i + c \sum_{i=n+1}^{2n} \xi_i + c \sum_{j=1}^{m} \zeta_j
$$

subject to

$$
\begin{aligned}
y^i(\theta \cdot x^i + \theta_0) &\geq 1 - \xi_i \quad \text{for all } i \in \{1, \ldots, 2n\} \\
v^j(\theta \cdot u^j + \theta_0) &\geq 1 - \zeta_j \quad \text{for all } j \in \{1, \ldots, m\} \\
\xi_i &\geq 0 \quad \text{for all } i \in \{1, \ldots, 2n\} \\
\zeta_j &\geq 0 \quad \text{for all } j \in \{1, \ldots, m\}
\end{aligned}
$$

which is essentially equal to Pat's suggestion. Therefore, Robin = Pat.

In summary, we have Pat $\neq$ Dana, Robin = Pat and Dana $\neq$ Robin.

4 Error bounds (15 points)

You have a data set of size n drawn from some data distribution P_D. You consider two hypothesis classes $\mathcal{H}_1 \subset \mathcal{H}_2$, with VC dimensions h_1 and h_2. You find f_1 that minimizes R_n on $\mathcal{H}_1$ with empirical risk er_1 and f_2 that minimizes R_n on $\mathcal{H}_2$ with empirical risk er_2.

4. (a) (5 points) Describe a theoretical method for selecting a hypothesis and under what circumstances it would select f_1.

Solution: Here we are trying to select a hypothesis from a nested set of hypothesis classes. We can apply the structural risk minimization (SRM) framework for this problem. In particular, we find and compare the bounds for the risk of f_i ($i = 1, 2$):

$$R(f_i) \leq er_i + \sqrt{\frac{h_i(\log(\frac{2n}{h_i}) + 1) + \log(\frac{4}{\delta})}{n}}$$

SRM selects the f_i with the lowest value of the bound. In particular, f_1 will be selected if its bound is less than that of f_2. Note that we used the bound that involves the VC dimension h_i. A common error was to use the bound for finite hypothesis classes.

(b) (5 points) Describe an empirical method for selecting a hypothesis and under what circumstances it would select f_1.

Solution: By 'empirical' we mean 'experimental' or 'in practice'. There are many reasonable and acceptable schemes to select a hypothesis. If we assumed access to more data, we can test hypothesis under new data and pick the one that performed the best (under 0-1 loss, or some other loss). If we did not have more data, we can also use cross validation in a number of ways. The essential operation is to split the given data into a training set and holdout set. We learn parameters with training set data, and evaluate the resulting hypotheses on the holdout set. This can either be used to decide the best hypothesis directly, or to decide the suitable hypothesis class and then choose the best hypothesis from within that class using all the given data. This can be extended to k-fold or leave-one-out CV if we repeat it on different partitions on the given data. f_1 is chosen if it performs the best on new/held out data.

Any similar approach is acceptable, as long as it was briefly explained. A common error was to say choose the hypothesis with lowest training error; this is not advisable in practice due to overfitting. Also, note that in our problem this always chooses f_2 (and hence f_1 only if they are the same hypothesis).

(c) (5 points) You have a competitor who was given a different data set of size n, also drawn from P_D. Imagine that you and your competitor are both limited to hypotheses in $\mathcal{H}_1$. Give a bound on how much better the generalization performance of your competitor's best hypothesis might be than your f_1, so that the bound will hold with probability at least 0.9.

Solution: We derive a bound similar to that in HW3 (Q11). Let our hypothesis be f_1, and the competitor's be f_c. Then:

$$R(f_1) - R(f_c) = [R(f_1) - R_n(f_1)] + [R_n(f_1) - R(f_c)]$$

So far the above holds regardless of which training set R_n is evaluated on. Ultimately this will not matter because the bounds we use work as long as the training sets come from the same distribution P_D, which is the case here. However, if we evaluate it on our data set, we can take advantage of the fact that f_1 is the empirical risk minimizer for our data set, so $R_n(f_1) \leq R_n(f_c)$. Applying this to the second $R_n(f_1)$ term above:

$$R(f_1) - R(f_c) \leq [R(f_1) - R_n(f_1)] + [R_n(f_c) - R(f_c)] \leq 2\sqrt{\frac{h_1(\log(\frac{2n}{h_1}) + 1) + \log(4/0.1)}{n}}$$

with probability at least $(1 - 0.1) = 0.9$. (We applied the usual bound twice.)
One common error was trying to apply the bounds directly to compare f_1 and f_c. This is incorrect because the bound only holds if the same hypothesis is involved. (The Chernoff bound applies when comparing an empirical estimate with its expectation.)