CS 4501: Algorithmic Economics
Assignment 3

Evan Zimmerman

Question 1

Consider two random variables X and Y with joint distribution $F(x,y)$: Prove the following two results:

- $E[X] = E_Y[E_X[X|Y]]$
- $Var(X) = E[Var_X(X|Y)] + Var_Y(E_X[X|Y])]$

Here $E_X[X|Y]$ is the conditional expectation of X given Y and $Var_X(X|Y)$ is the conditional variance.

$E[X] = E_Y[E_X[X|Y]]$ Proof:

$$
\begin{aligned}
E_Y[E_X[X|Y]] &= \int\limits_{supp(Y)} E_X[X|Y]f(y)\,dy && (1)\\
&= \int\limits_{supp(Y)}\int\limits_{supp(X)} xf_{X|Y}(x|y)\,dxf(y)\,dy && (2)\\
&= \int\limits_{supp(Y)}\int\limits_{supp(X)} x\frac{f(x,y)}{f(y)}\,dxf(y)\,dy && (3)\\
&= \int\limits_{supp(Y)}\int\limits_{supp(X)} x\frac{f(x,y)}{f(y)}f(y)\,dx\,dy && (4)\\
&= \int\limits_{supp(Y)}\int\limits_{supp(X)} xf(x,y)\,dx\,dy && (5)\\
\text{by fubini's theorem} & && (6)\\
&= \int\limits_{supp(X)}\int\limits_{supp(Y)} xf(x,y)\,dy\,dx && (7)\\
&= \int\limits_{supp(X)} x\int\limits_{supp(Y)} f(x,y)\,dy\,dx && (8)\\
&= \int\limits_{supp(X)} xf(x)\,dx && (9)\\
&= E[X] && (10)
\end{aligned}
$$

$Var(X) = E[Var_X(X|Y)] + Var_Y(E_X[X|Y])]$ Proof:

$$Var(X) = E[X^2] - E[X]^2 \tag{11}$$
$$E[X^2] = Var(X) + E[X]^2 \tag{12}$$
$$\text{by the law of total expectation} \tag{13}$$
$$= E[Var_X(X|Y) + E[X|Y]^2] \tag{14}$$
$$E[X^2] - E[X]^2 = E[Var_X(X|Y) + E[X|Y]^2] - E[X]^2 \tag{15}$$
$$\text{by the law of total expectation on } E[X]^2 \text{ ,} \tag{16}$$
$$E[X^2] - E[X]^2 = E[Var_X(X|Y) + E[X|Y]^2] - E[E[X|Y]]^2 \tag{17}$$
$$E[X^2] - E[X]^2 = E[Var_X(X|Y)] + E[E[X|Y]^2]] - E[E[X|Y]]^2 \tag{18}$$
$$Var(X) = E[Var_X(X|Y)] + Var_Y(E[X|Y]) \tag{19}$$

Question 2

Denote the loss matrix for a multi-class classification problem as $\mathcal{L}$, where $\mathcal{L}_{kj}$ suggests the loss for classifying the true class $\mathcal{C}_k$ as class $\mathcal{C}_j$. For a given input vector x, our uncertainty in the true class $\mathcal{C}_k$ is expressed through the joint probability distribution $p(x, \mathcal{C}_k)$.

Write down your decision rule for classifying x. (5 pts, hint: expected loss minimization)

$$\text{Classify } x \text{ as class } \mathcal{C}_j \text{ where } j = \arg\min_j \sum_k \mathcal{L}_{kj}\, p(\mathcal{C}_k|x). \tag{20}$$

Now impose a special structure on $\mathcal{L}$: $\mathcal{L}_{kj}$ = $1 - \mathcal{I}_{kj}$ where $\mathcal{I}$ is an identity matrix. How does this special structure on L simplify your decision rule for classifying x? (15 pts, hint: consider what matters more for expected loss minimization)

This special structure means that if we make the correct classification $(k = j)$ then $\mathcal{L}_{kj} = 0$ (no loss), but if we make an incorrect classification $(k \neq j)$ then $\mathcal{L}_{kj} = 1$. This special structure simplifies our decision rule for classifying x. With this special structure the expected loss of classifying x as $\mathcal{C}_j$ is greatly simplified:

$$E[L|\mathcal{C}_j] = \sum_k (1 - \mathcal{I}_{kj})\, p(\mathcal{C}_k|x) \tag{21}$$
$$= \sum_{k \neq j} p(\mathcal{C}_k|x) \tag{22}$$
$$= 1 - \; p(\mathcal{C}_j|x) \tag{23}$$

So the simplified decision rule becomes:

$$\text{Classify } x \text{ as class } \mathcal{C}_j \text{ where} j = \arg\max_j p(\mathcal{C}_j|x). \tag{24}$$

Finally you are given a rejection option, i.e., you can choose not to predict x's class but to incur loss λ. Find the decision criterion based on the selection of a rejection threshold θ for data x that will give the minimum expected loss under general structure of $\mathcal{L}$ and the special structure of $\mathcal{L}$ mentioned above. Whats the relationship between θ and λ (20 pts).

General structure of $\mathcal{L}$:
Classify x as $\mathcal{C}_j$ if the expected loss for $\mathcal{C}_j$ is the minimum and is less than λ. If the expected loss is greater than λ, choose not to predict x's class.

$$\text{If } \arg\min_j \sum_k \mathcal{L}_{kj}\, p(\mathcal{C}_k|x) < \lambda, \text{ follow decision rule: } \arg\min_j \sum_k \mathcal{L}_{kj}\, p(\mathcal{C}_k|x). \text{ Otherwise, do not classify } x \text{ and incur loss } \lambda. \tag{25}$$

Special Structure $\mathcal{L}_{kj} = 1 - \mathcal{I}_{kj}$:
Classify x as $\mathcal{C}_j$ if $p(\mathcal{C}_j|x)$ is greater than the rejection threshold θ that is determined by the rejection loss λ.

$$\text{If } \arg\max_j p(\mathcal{C}_j|x) > \theta, \text{ follow decision rule: } \arg\max_j p(\mathcal{C}_j|x). \text{ Otherwise, do not classify } x \text{ and incur loss } \lambda. \tag{26}$$

The relationship between the rejection threshold θ and the rejection loss λ comes from the condition that expected loss of classifying x as $\mathcal{C}_j$ should be less than λ.

$$1 - p(\mathcal{C}_j|x) < \lambda \tag{27}$$
$$p(\mathcal{C}_j|x) > 1 - \lambda \tag{28}$$
$$\text{So the rejection threshold } \theta = 1 - \lambda \tag{29}$$

Question 3

Given two hypotheses h_1 and h_2, we define h = $h_1 \cap h_2$ as a new hypothesis that labels an example $+1$ only if both h_1 and h_2 label it as $+1$, otherwise -1. We can extend this concept to sets of hypotheses: given two sets of hypotheses H_1 and H_2, define $H^* = \{h_1 \cap h_2; h_1 \in H_1; h_2 \in H_2\}$. Suppose the shattering coefficient of H_1 is $H_1[n]$ (i.e., the maximum number of ways that the hypothesis class H_1 can label a set of n points is $H_1[n]$). Similarly, suppose that the shattering coefficient of H_2 is $H_2[n]$. Prove that $H^* \leq H_1[n]H_2[n]$.

Consider a set S of n points. The maximum number of ways that the hypothesis class H_1 can label S is $H_1[n]$ and the maximum number of ways that the hypothesis class H_2 can label S is $H_2[n]$. For each labeling done by a hypothesis $h_1 \in H_1$ there are at most $H_2[n]$ ways to label S using hypotheses from H_2. If a point is labeled as $+1$ by h_1 a hypothesis $h_2 \in H_2$ will label it the same or as -1. Points labeled as -1 by either h_1 or h_2 will always be labeled as -1 from the combined hypothesis h = $h_1 \cap h_2$. Therefore for each of the $H_1[n]$ labelings from H_1 there are at most $H_2[n]$ compatible labelings from H_2. Thus, $H_1[n]H_2[n]$ is the maximum number of distinct labelings that could be produced by H^*. This applies to any set S of n points. This argument proves $H^* \leq H_1[n]H_2[n]$.