CS475 Machine Learning, Fall 2012: Homework 2

Daniel Crankshaw - dcranks1

September 26, 2012

Question 1:

(a) $P = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$

(b) Let H be a random variable representing the number of heads flipped in 10 coin tosses. $P(H \geq 3) = \sum_{i=3}^{10} P(H = i) = \sum_{i=3}^{10} \binom{10}{i} = 968.$

(c) It takes 14 flips on average.

Question 2:

(a)

$$\mathbb{P}(X) = \sum_{j=1}^{k} \mathbb{P}(X|C = j)\mathbb{P}(C = j) \quad (1)$$

(b)

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}[\mathbb{E}(X|C)] && (2) \\ &= \sum_{j=1}^{k} \mathbb{E}(X|C = j)\mathbb{P}(C = j) && (3) \\ &= \sum_{j=1}^{k} \mu_j \pi_j && (4) \end{aligned}$$

(c)

$$\begin{aligned} log(\mathbb{P}D|\mu_j, \sigma_j for j = 1, \ldots, k) &= \sum_{i=1}^{n} log(\mathbb{P}(x_n|\mu_j, \sigma_j)) && (5) \\ & && (6) \\ &= \sum_{i=1}^{n} log\left[\sum_{j=1}^{k} \pi_j \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right\}\right] && (7) \\ & && (8) \end{aligned}$$

Question 3: First I will prove the first part of the set of inequalities, and then the second part. For

$$\|y - X_1\hat{\beta}_3\|_2^2 \geq \|y - X_1\hat{\beta}_0\|_2^2 \tag{9}$$

Assume that $\|y - X_1\hat{\beta}_3\|_2^2 < \|y - X_1\hat{\beta}_0\|_2^2$. Then there exists some $\beta_0' = \hat{\beta}_3$ such that $\|y - X_1\beta_0'\|_2^2 < \|y - X_1\hat{\beta}_0\|_2^2$. Thus, $\hat{\beta}_0$ is not the argmin. But $\hat{\beta}_0$ is the argmin, and so we have reached a contradiction. Therefore, $\|y - X_1\hat{\beta}_3\|_2^2 \geq \|y - X_1\hat{\beta}_0\|_2^2$.

$$\|y - X_1\hat{\beta}_0\|_2^2 \geq \|y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\|_2^2 \tag{10}$$

Notice that $\text{argmax}_{\beta_0} X_1\beta_0 \geq \mathbf{0}$ (where $\mathbf{0}$ is the n-dimensional null vector). Similarly, $\text{argmax}_{\beta_1} X_1\beta_1 \geq \mathbf{0}$ and $\text{argmax}_{\beta_2} X_2\beta_2 \geq \mathbf{0}$. And because we are subtracting each of these quantities from y, no matter what $X_1\hat{\beta}_0$ is, we can always get the same value from $X_1\hat{\beta}_1$. And because $\text{argmax}_{\beta_2} X_2\hat{\beta}_2 \geq \mathbf{0}$ then we either have $X_2\hat{\beta}_2 = \mathbf{0}$ in which case the two expressions are equal, or $X_2\hat{\beta}_2 > \mathbf{0}$ in which case the $\|y - X_1\hat{\beta}_0\|_2^2 > \|y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\|_2^2$. Thus the inequality holds true.

Question 4:

(a) With non-overlapping samples, in the worst case every split exactly divides the data in half. If we divide the data in half at every split, then each sample needs $log_2 n$ splits to be labeled. Because the samples are non-overlapping, there will never be a split that contains data with the same set of features with different labels. Thus, we can always perform a series of feature splits that will result in a node in the decision tree with all the labels agreeing. Therefore, we can perform a perfect labeling. Therefore, in this situation, we can always perform a perfect labeling of all n samples with a decision tree of depth at most $log_2 n$.

(b) We cannot. If we lose the assumption that all labels are non-overlapping, then we can have a situation where we have at least two samples with the exact same feature vector that have different labels. This situation means that even after $log_2 n$ feature splits (isolating all samples with this feature vector from any with distinct feature vectors) we will have a set of data at a node which do not all have the same label. Therefore, we cannot perfectly label the data.

Question 5:

(a) Convex formulations have exactly one minimum, meaning that the global minimum is also the only local minimum. This makes it much easier to find the minimum of the function.

(b) The definition of convexity is as follows. A function $f(x)$ is convex if for any two points x_1, x_2 in X and $t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) \tag{11}$$

$$
\begin{aligned}
h(x) &= f(x) + g(x) && (12)\\
h(tx_1 + (1-t)x_2) &= f(tx_1 + (1-t)x_2) + g(tx_1 + (1-t)x_2) && (13)\\
h(tx_1 + (1-t)x_2) &\leq tf(x_1) + tg(x_1) + (1-t)f(x_2) + (1-t)g(x)2 \quad \text{(by the definition of convexity)} && (14)\\
&= t(f(x_1) + g(x_1)) + (1-t)(f(x_2) + g(x_2)) && (15)\\
&= th(x_1) + (1-t)h(x_2) && (16)
\end{aligned}
$$

Therefore, $h(tx_1+(1-t)x_2) \leq th(x_1)+(1-t)h(x_2)$ and thus satisfies the definition of a convex function.

(c) $f'(x) = x^4$ and $g'(x) = x^2$ results in the a non-convex function $f'(x) - g'(x)$.
$f(x) = e^x$ and $g(x) = 2^x$ results in a convex function $f(x) - g(x)$.