10-701 Machine Learning Fall 2011: Homework 2 Solutions

1 Linear regression, model selection

1.1 Ridge regression

Starting from our true model $\mathbf{y} = \mathbf{X}\theta + \epsilon$, we express $\hat{\theta}$ in terms of ϵ and θ:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\theta + \epsilon \\ \left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\mathbf{y} &= \left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\left(\mathbf{X}\theta + \epsilon\right) \\ \hat{\theta} &= \left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\mathbf{X}\theta + \left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\epsilon. \end{aligned}$$

Because $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$, it follows from the hint that

$$\begin{aligned} \left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\epsilon &\sim \mathcal{N}\left(0, \sigma^2\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\left(\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\right)^\top\right) \\ &= \mathcal{N}\left(0, \sigma^2\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\right). \end{aligned}$$

Hence

$$\hat{\theta} \sim \mathcal{N}\left(\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\mathbf{X}\theta,\ \sigma^2\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\right),$$

in other words $\hat{\theta}$ has a Gaussian distribution with mean and covariance

$$\begin{aligned} E\left[\hat{\theta}\right] = \mu &= \left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\mathbf{X}\theta \neq \theta \\ \Sigma &= \sigma^2\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}, \end{aligned}$$

implying that ridge regression is biased.

1.2 Extra features

Let β_{new} be the parameter corresponding to the $n \times 1$ vector of new features X_{new}, and define

$$B = \begin{bmatrix} \beta \\ \beta_{new} \end{bmatrix}$$

to be the original vector β concatenated with β_{new}. Also, let $\mathrm{J}_{new}(B) = \left(\mathbf{X}_{new}B - y\right)^\top\left(\mathbf{X}_{new}B - y\right)$ be the squared error for the augmented feature values $\mathbf{X}_{new}$.

Let's derive a few facts beforehand. The assumption $\mathbf{X}_{new}^\top\mathbf{X}_{new} = I$ implies that the columns of $\mathbf{X}_{new}$ are orthonormal, which in turn implies:

1. $\mathbf{X}^\top\mathbf{X} = I$
2. $\mathbf{X}^\top X_{new} = \vec{0}$ where $\vec{0}$ is the zero vector

3. $X_{new}^{\top} X = 1$

We know that the minimizer for $\mathrm{J}_{new}(B)$ is

$$\begin{aligned} \hat{B} &= \left(\mathbf{X}_{new}^{\top}\mathbf{X}_{new}\right)^{-1}\mathbf{X}_{new}^{\top}\mathbf{y} \\ &= \mathbf{X}_{new}^{\top}\mathbf{y} \quad (\text{because } \mathbf{X}_{new}^{\top}\mathbf{X}_{new} = I), \end{aligned}$$

while the minimizer for the original problem $\mathrm{J}_1(\beta)$ is

$$\begin{aligned} \hat{\beta} &= \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{y} \\ &= \mathbf{X}^{\top}\mathbf{y} \quad (\text{because } \mathbf{X}^{\top}\mathbf{X} = I). \end{aligned}$$

Given the above facts, the minimum value of $\mathrm{J}_{new}(B)$ is given by

$$\begin{aligned} \mathrm{J}_{new}\left(\hat{B}\right) &= \left(\mathbf{X}_{new}\hat{B} - \mathbf{y}\right)^{\top}\left(\mathbf{X}_{new}\hat{B} - \mathbf{y}\right) \\ &= \left(\mathbf{X}_{new}\mathbf{X}_{new}^{\top}\mathbf{y} - \mathbf{y}\right)^{\top}\left(\mathbf{X}_{new}\mathbf{X}_{new}^{\top}\mathbf{y} - \mathbf{y}\right) \\ &= \left(\left(\mathbf{X}\mathbf{X}^{\top} + X_{new}X_{new}^{\top}\right)\mathbf{y} - \mathbf{y}\right)^{\top}\left(\left(\mathbf{X}\mathbf{X}^{\top} + X_{new}X_{new}^{\top}\right)\mathbf{y} - \mathbf{y}\right) \\ &= \left(\left(\mathbf{X}\mathbf{X}^{\top}\mathbf{y} - \mathbf{y}\right) + X_{new}X_{new}^{\top}\mathbf{y}\right)^{\top}\left(\left(\mathbf{X}\mathbf{X}^{\top}\mathbf{y} - \mathbf{y}\right) + X_{new}X_{new}^{\top}\mathbf{y}\right) \\ &= \left(\mathbf{X}\mathbf{X}^{\top}\mathbf{y} - \mathbf{y}\right)^{\top}\left(\mathbf{X}\mathbf{X}^{\top}\mathbf{y} - \mathbf{y}\right) + 2\left(\mathbf{X}\mathbf{X}^{\top}\mathbf{y} - \mathbf{y}\right)^{\top}\left(X_{new}X_{new}^{\top}\mathbf{y}\right) + \left(X_{new}X_{new}^{\top}\mathbf{y}\right)^{\top}\left(X_{new}X_{new}^{\top}\mathbf{y}\right). \end{aligned}$$

Observing that $\mathrm{J}_1\left(\hat{\beta}\right) = \left(\mathbf{X}\mathbf{X}^{\top}\mathbf{y} - \mathbf{y}\right)^{\top}\left(\mathbf{X}\mathbf{X}^{\top}\mathbf{y} - \mathbf{y}\right)$, we get

$$\begin{aligned} \mathrm{J}_{new}\left(\hat{B}\right) &= \mathrm{J}_1\left(\hat{\beta}\right) + 2\left(\mathbf{X}\mathbf{X}^{\top}\mathbf{y} - \mathbf{y}\right)^{\top}\left(X_{new}X_{new}^{\top}\mathbf{y}\right) + \left(X_{new}X_{new}^{\top}\mathbf{y}\right)^{\top}\left(X_{new}X_{new}^{\top}\mathbf{y}\right) \\ &= \mathrm{J}_1\left(\hat{\beta}\right) + 2\left(\mathbf{X}\mathbf{X}^{\top}\mathbf{y}\right)^{\top}\left(X_{new}X_{new}^{\top}\mathbf{y}\right) - 2\mathbf{y}^{\top}\left(X_{new}X_{new}^{\top}\mathbf{y}\right) + \left(X_{new}X_{new}^{\top}\mathbf{y}\right)^{\top}\left(X_{new}X_{new}^{\top}\mathbf{y}\right) \\ &= \mathrm{J}_1\left(\hat{\beta}\right) + 2\left(\mathbf{y}^{\top}\mathbf{X}\mathbf{X}^{\top}X_{new}X_{new}^{\top}\mathbf{y}\right) - 2\left(\mathbf{y}^{\top}X_{new}X_{new}^{\top}\mathbf{y}\right) + \left(\mathbf{y}^{\top}X_{new}X_{new}^{\top}X_{new}X_{new}^{\top}\mathbf{y}\right) \\ &= \mathrm{J}_1\left(\hat{\beta}\right) + 2\left(\mathbf{y}^{\top}\mathbf{X}\vec{0}X_{new}^{\top}\mathbf{y}\right) - 2\left(\mathbf{y}^{\top}X_{new}X_{new}^{\top}\mathbf{y}\right) + \left(\mathbf{y}^{\top}X_{new}X_{new}^{\top}\mathbf{y}\right) \\ &= \mathrm{J}_1\left(\hat{\beta}\right) - \mathbf{y}^{\top}X_{new}X_{new}^{\top}\mathbf{y} \\ &= \mathrm{J}_1\left(\hat{\beta}\right) - \left(X_{new}^{\top}\mathbf{y}\right)^{\top}\left(X_{new}^{\top}\mathbf{y}\right) \\ &\leq \mathrm{J}_1\left(\hat{\beta}\right). \end{aligned}$$

In other words, our new feature X_{new} allows us to decrease (or at least maintain) the minimized objective value.

1.3 Model Complexity

The above result says that we can further decrease (or at least maintain) the squared error objective by adding new features. However, notice that this error objective is computed on the training samples, and not the true data distribution. Reducing the training error does not guarantee a reduction in error on the true distribution; in fact, practice suggests that reducing the training error beyond a certain point actually increases error on the true data distribution (as represented by test samples). In conclusion, increasing features does not guarantee a better model (with respect to the true data distribution).

1.4 Overfitting

To address overfitting, we need to pick the candidate model with the lowest error on the true data distribution. We can do so by performing leave-one-out-cross-validation (LOOCV), in which we train each candidate model

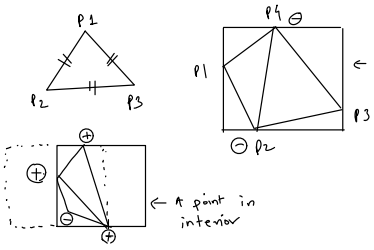


Figure 1: Shattering points with squares

on $n-1$ data points, then compute its error on the remaining "test" data point. We repeat this for all n choices of the test data point, and average the test data point errors to get the LOOCV error for the candidate model. The LOOCV error approximates the error on the true data distribution, hence the best candidate model is the one with lowest LOOCV.

2 Learning Theory

2.1 PAC learning

True. PAC learning only assures that with probability $1-\delta$, the error of the returned hypothesis is less than ϵ.

2.2 VC Dimension

1. $\mathcal{X} = \mathbb{R}$, $\mathcal{H}$ is the union of 2 intervals: VC Dimension 4.

   A union of two intervals allow us to correctly label a point set of the form $\ominus \oplus \ominus \oplus \ominus$. All labelings for 4 points can be easily shown to be consistent with this label format. For 5 points, however, there exists the labeling $\oplus \ominus \oplus \ominus \oplus$ which cannot be accomplished by the union of two intervals.

2. $\mathcal{X} = \mathbb{R}^2$, $\mathcal{H}$ is the set of axis-parallel squares (with equal height and width) :VC dimension 3

   As shown in Figure 1, a square can shatter the 3 points that are the vertices of an equilateral triangle.

   Consider the bounding square (the smallest square so that all points lie within or on the square) for any 4 points. By moving the square along the x or y axis, the label at an individual point can be changed.

   We will assume that the 4 points form a convex shape - if not, at least one point lies in the convex hull of the 4 points and a labeling where the interior point is labeled -1 and the external ones are labeled +1 cannot be achieved.

   The tightest bounding square must touch at least 3 points out of 4 (if not, we can make it smaller until this happens). Since each diagonal of the quadrilateral has 2 vertices, at least one of the two diagonals must have both points on the sides of the square. It can be shown then, for instance, that if both the points on that diagonals are labeled the same, then of the 4 possible labelings for the remaining 2 points, at least one of cannot be attained.

   (Note: This is not a completely rigorous proof. As the figure shows by a dotted line, the bounding square is not unique and an exhaustive and rigorous proof would need to account for that.)

3. $\mathcal{X} = \mathbb{R}^3$, $\mathcal{H}$ is the set of axis-parallel rectangles: VC dimension 6.

   Consider the 6 points (1,0,0),(0,1,0),(0,0,1),(1,0,0),(0,1,0),(0,0,1). If we draw a bounding box for these points, then by excluding/including each point by moving a face of the box, we can get any labeling for the points. So the VC dimension is at least 6.

   For 7 points, consider the bounding box. If the bounding rectangle has at least one point in its interior, then we cannot accomplish the labeling where the interior point is labeled $\ominus$ and the rest are labeled

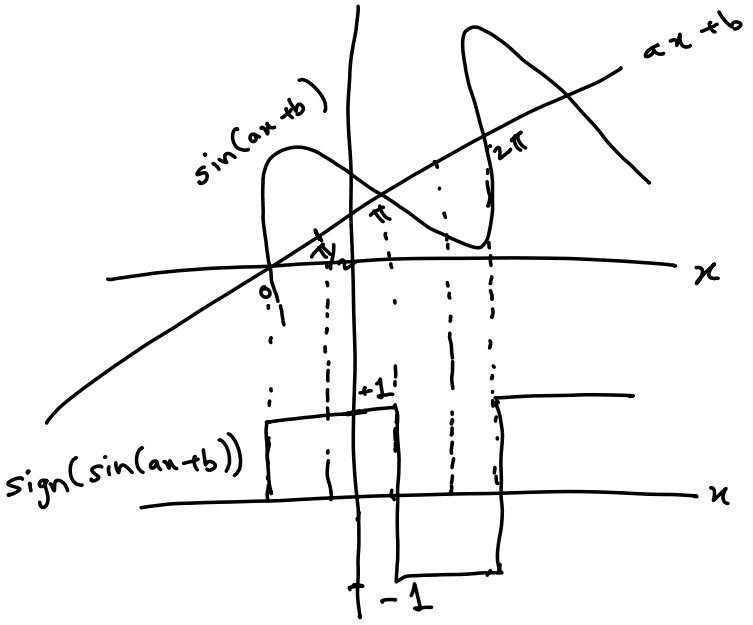


Figure 2: Functional depiction of $\text{sign}(\sin(ax+b))$.

$\oplus$. If none of the points are in the interior, then at least two must be on the same face of the rectangle by the piegonhole principle and then cannot have opposite labels.

4. $\mathcal{X} = \mathbb{R}$, $\mathcal{H}$ is the set of functions defined as following:

$$h(x) = \text{sign}(\sin(ax+b)) \tag{1}$$

where a and b are parameters in h, $\text{sign}(x) = +1$ if $x \geq 0$ and $\text{sign}(x) = -1$ otherwise. : VC dimension ∞.

From Figure 2 we can see that $h(x)$ is a square wave function with amplitude 1 and period given by $2\pi \cos\tan^{-1}(a)$.

[Non-rigorous argument - Therefore, by changing the value of a and b, the square wave frequency can be manipulated to produce any labeling for a given set of points. Thus, its VC dimension is infinite.]

Construction given in slides: It is enough to show that for any l, the set of points $\{x_1, \cdots, x_l\}$ can be shattered. Consider the set of points given by

$$x_i = 10^{-i} \tag{2}$$

Choose any labeling $\{y_1, \cdots, y_l\}$. Then let

$$a = \pi \left(1 + \sum_{i=1}^{l} \frac{(1-y_i)10^i}{2}\right) \tag{3}$$

and $b = 0$.

Then

$$h(x_j) = \text{sign}\left(\sin\left(10^{-j} \times \pi \left(1 + \sum_{i=1}^{l} \frac{(1-y_i)10^i}{2}\right)\right)\right) \tag{4}$$

$$= \text{sign}\left(\sin\left(10^{-j}\pi + \sum_{i=1}^{l}(1-y_i)10^{i-j}\frac{\pi}{2}\right)\right) \tag{5}$$

For any $y_i = 1$, the corresponding term in the summation will be zero. Also, any $i > j$, we will be adding an integral number of $\frac{\pi}{2}$ terms to a sine function, which causes no change in value. So those terms can be dropped from the summation.

$$
\begin{aligned}
h(x_j) &= \text{sign}\left(\sin\left(10^{-j}\pi + \sum_{i:i\leq j, y_i=-1}(1-y_i)10^{i-j}\frac{\pi}{2}\right)\right) && (6)\\
&= \text{sign}\left(\sin\left(10^{-j}\pi + (1-y_j)\frac{\pi}{2} + \sum_{i:i<j, y_i=-1}2\times 10^{i-j}\frac{\pi}{2}\right)\right) && (7)\\
&= \text{sign}\left(\sin\left((1-y_j)\frac{\pi}{2} + 10^{-j}\pi + \pi\sum_{i:i<j, y_i=-1}\times 10^{i-j}\right)\right) && (8)\\
& && (9)
\end{aligned}
$$

Where we use the fact that $\sin(\pi + x) = -\sin(x)$. It is easy to show that the summation in the last term (and the second term) is always less than 1. Therefore, if $y_j = 1$, the argument of the sine function is between 0 and π. Therefore the sine function takes positive values and $h(x_j) = 1 = y_j$. If $y_j = -1$, the first term becomes π and the argument to the sine function is between π and 2π. The sine function takes a negative value and $h(x_j) = -1 = y_j$.

Thus $h(x_j) = y_j$ for all j. So the set $\{10^{-1}, \cdots, 10^{-l}\}$ can be shattered for any value for l. Therefore the VC dimension is infinite.

5. $\mathcal{H}$ is a finite hypothesis class such that $|\mathcal{H}| < \infty$. Show that the VC dimension of $\mathcal{H}$ is upper bounded by $\log_2 |\mathcal{H}|$. (Hint: how many different labelings are there for a set of size $\log_2 |\mathcal{H}| + 1$?)

   Consider a set of size $\log_2 |\mathcal{H}| + 1$. Since each point can belong to one of two classes, the set can be labeled in $2^{(\log_2 |\mathcal{H}|+1)} = 2|\mathcal{H}|$ ways.

   For a given set of points, a particular hypothesis provides a single labeling. Therefore $|\mathcal{H}|$ can at most only provide $|\mathcal{H}|$ labelings for the chosen set. So it cannot shatter a set of size $\log_2 |\mathcal{H}| + 1$.

   Therefore its VC dimension is bounded above by $\log_2 |\mathcal{H}| + 1$

2.3 The Approximation-Estimation-Optimization Tradeoff

1. Filling in the table:

Table 1: Typical variations when $\mathcal{F}$, n and ρ increase.

	$\mathcal{F}$	n	ρ
ε_{app}	$\downarrow$	$\times$	$\times$
ε_{est}	$\uparrow$	$\downarrow$	$\times$
ε_{opt}	$\times$	$\times$	$\uparrow$
T	$\uparrow$	$\uparrow$	$\downarrow$

Effect on ε_{app}: As $\mathcal{F}$ becomes larger, $\mathbb{E}[R(f^*_{\mathcal{F}})]$, the risk of the best choice of f from $\mathcal{F}$ decreases. So ε_{app} decreases. As n grows larger, there is no effect because the true risk expressions do not depend on n.

Effect on ε_{est}: As $\mathcal{F}$ becomes larger, $\mathbb{E}[R(f^*_{\mathcal{F}})]$, the risk of the best choice of f from $\mathcal{F}$ decreases. So ε_{est} increases due to its definition. As n grows larger, the error decreases.

Effect on ε_{opt}: As ρ grows larger, the tolerance grows larger and a less accurate solution is acceptable. So the error increases.

Effect on T: As $\mathcal{F}$ becomes larger, the search space increases so the time required increases. As n increases, the computation of the empirical risk becomes slower, so T increases. As ρ increases,a less accurate solution is acceptable, so the time for search decreases.

2. It is not necessary to accurately solve the minimization $f_n = \arg\min_{f \in \mathcal{F}} R_n(f)$ because if the ε_{opt} term is smaller than the other terms for more than once choice of f_n, then many different choices of f_n will still give us nearly the same overall error.

2.4 The Approximation-Estimation-Optimization Tradeoff

1. Filling in the table:

Table 2: Typical variations when $\mathcal{F}$, n and ρ increase.

	$\mathcal{F}$	n	ρ
ε_{app}	$\downarrow$	$\times$	$\times$
ε_{est}	$\uparrow$	$\downarrow$	$\times$
ε_{opt}	$\times$	$\times$	$\uparrow$
T	$\uparrow$	$\uparrow$	$\downarrow$

Effect on ε_{app}: As $\mathcal{F}$ becomes larger, $\mathbb{E}[R(f^*_{\mathcal{F}})]$, the risk of the best choice of f from $\mathcal{F}$ decreases. So ε_{app} decreases. As n grows larger, there is no effect because the true risk expressions do not depend on n.

Effect on ε_{est}: As $\mathcal{F}$ becomes larger, $\mathbb{E}[R(f^*_{\mathcal{F}})]$, the risk of the best choice of f from $\mathcal{F}$ decreases. So ε_{est} increases due to its definition. As n grows larger, the error decreases.

Effect on ε_{opt}: As ρ grows larger, the tolerance grows larger and a less accurate solution is acceptable. So the error increases.

Effect on T: As $\mathcal{F}$ becomes larger, the search space increases so the time required increases. As n increases, the computation of the empirical risk becomes slower, so T increases. As ρ increases,a less accurate solution is acceptable, so the time for search decreases.

2. It is not necessary to accurately solve the minimization $f_n = \arg\min_{f \in \mathcal{F}} R_n(f)$ because if the ε_{opt} term is smaller than the other terms for more than once choice of f_n, then many different choices of f_n will still give us nearly the same overall error.

3 Expectation-Maximization algorithm [Suyash, 25 points]

In this question, we will derive some details about the expectation-maximization (EM) algorithm and work out an example application.

3.1 True-False Questions [6 points]

Explain whether the following statements are true or false with a single line of explanation.

1. **(2 points)** The EM algorithm maximizes the complete log-likelihood.

   Answer: False. The EM algorithm is coordinate-ascent on the free energy functional, which is a lower bound on the incomplete likelihood.

2. **(2 points)** Even if the complete log-likelihood is multi-modal, EM cannot get stuck in a local minima.

   Answer: False. EM can stuck in local minima.

3. **(2 points)** The free-energy functional that EM optimizes is a lower bound on the complete log-likelihood.

   Answer: False.The free-energy functional is a lower bound on the incomplete likelihood.

3.2 EM and KL divergence [7 points]

3.2.1 Optimizing the free-energy functional [4 points]

Consider the free-energy functional discussed in class that the EM algorithm optimizes. The free-energy functional is given by:

$$F(q,\theta) = \sum_{q(z|x)} q(z|x) \log \frac{p(z,x|\theta)}{q(z|x)} \tag{10}$$

for any probability distribution $q(z)$. (Note: In this setting, the only random variables are the z variables). In class, we claimed that

$$\log p(x|\theta) - F(q,\theta) = KL(q \,\|\, p(z|x,\theta)) \tag{11}$$

where $KL(a \,\|\, b) = \sum_x a(x) \log \frac{a(x)}{b(x)}$ is called the Kullback-Leibler divergence between two probability distributions a and b. Show that this result is true.

Proof:

Since

$$\begin{aligned}
F(q,\theta) &= \sum_{q(z|x)} q(z|x) \log \frac{p(z,x|\theta)}{q(z|x)} \\
&= \sum_{q(z|x)} q(z|x) \log \frac{p(z|x,\theta)p(x|\theta)}{q(z|x)} \\
&= \sum_{q(z|x)} q(z|x) \log \frac{p(z|x,\theta)}{q(z|x)} + \sum_{q(z|x)} q(z|x) \log p(x|\theta) \\
&= \sum_{q(z|x)} q(z|x) \log \frac{p(z|x,\theta)}{q(z|x)} + \log p(x|\theta)
\end{aligned}$$

Hence,

$$\begin{aligned}
\log p(x|\theta) - F(q,\theta) &= \log p(x|\theta) - \Big(\sum_{q(z|x)} q(z|x) \log \frac{p(z|x,\theta)}{q(z|x)} + \log p(x|\theta) \Big) \\
&= - \sum_{q(z|x)} q(z|x) \log \frac{p(z|x,\theta)}{q(z|x)} \\
&= KL(q \,\|\, p(z|x,\theta))
\end{aligned}$$

3.2.2 Optimal value of q(z) in EM [3 points]

Using the result in Equation 11, find the value for $q(z)$ that maximizes $F(q,\theta)$. (Hint: $KL(a \,\|\, b) \geq 0$, with equality if and only if $a = b$.)

Answer: To maximize $F(q,\theta)$, we need to minimize $KL(a \,\|\, b)$. Since $KL(a \,\|\, b) \geq 0$, with equality if and only if $a = b$. The value for $q(z)$ that maximizes $F(q,\theta)$ is $p(z|x,\theta)$.

3.3 Example of coin-tossing [12 points]

Consider the scenario where a sequence of heads and tails are being generated by tossing two coins independently. For each toss, the first coin is chosen with probability π and the second coin is chosen with probability $1-\pi$. The head probabilities of the two coins are given by p_1 and p_2. The probabilities π, p_1 and p_2 are unknown to us and we want to find their values.

Let us represent the result of the N coin tosses by $x_1, x_2, \cdots, x_N$. $x_i = 1$ indicates that that i^{th} toss came out heads and 0 otherwise. Let $z_1, z_2, \cdots, z_N$ indicate the coin which was tossed each time. For ease of mathematical analysis, suppose that $z_i = 1$ means the first coin was used for the i^{th} toss and $z_i = 0$ means the second coin was used for the i^{th} toss.

3.3.1 All variables observed

Suppose that we are allowed to observe which coin was used for each toss, i.e, the values of the variables $z_1, \cdots, z_N$ are known. The expression for the complete log-likelihood of the N coin tosses is

$$\begin{aligned} l_c(X, Z; \pi, p_1, p_2) &= \log \prod_{i=1}^{N} p(x_i, z_i; \pi, p_1, p_2) \\ &= \sum_{i=1}^{N} \log \left[\pi p_1^{x_i}(1-p_1)^{1-x_i}\right]^{z_i} \left[(1-\pi) p_2^{x_i}(1-p_2)^{1-x_i}\right]^{1-z_i} \\ &= \sum_{i=1}^{N} [z_i(\log \pi + x_i \log p_1 + (1-x_i)\log(1-p_1)) \\ &\qquad + (1-z_i)(\log(1-\pi) + x_i \log p_2 + (1-x_i)\log(1-p_2))] \end{aligned}$$

and the maximum likelihood equations for the parameters in this setting are:

$$\pi = \frac{\sum_{i=1}^{N} z_i}{N} \tag{12}$$

$$p_1 = \frac{\sum_{i=1}^{N} z_i x_i}{\sum_{i=1}^{N} z_i} \tag{13}$$

$$p_2 = \frac{\sum_{i=1}^{N} (1-z_i) x_i}{\sum_{i=1}^{N} (1-z_i)}. \tag{14}$$

3.3.2 Only x variables observed

Suppose we are not allowed to observe which coin was used for each toss, i.e, the value of the variables $z_1, \cdots, z_N$ are unknown. We will now use the EM algorithm for finding the values of the unknown parameters.

1. **(3 points)** Write the expression for the incomplete log-likelihood.

   Answer:

$$\begin{aligned} l_c(X; \pi, p_1, p_2) &= \log \prod_{i=1}^{N} p(x_i; \pi, p_1, p_2) \\ &= \log \prod_{i=1}^{N} \sum_{z_i} p(x_i, z_i; \pi, p_1, p_2) \\ &= \sum_{i=1}^{N} \log \sum_{z_i} \left[\pi p_1^{x_i}(1-p_1)^{1-x_i}\right]^{z_i} \left[(1-\pi) p_2^{x_i}(1-p_2)^{1-x_i}\right]^{1-z_i} \end{aligned}$$

2. **(5 points)** The E-step for our EM algorithm requires us to compute $q^*(z) = p(z|x; \pi^{(t)}, p_1^{(t)}, p_2^{(t)})$ where the superscript t indicates the value of the parameters at step t. What is the expression for the conditional probability $p(z_i = 1|x_i; \pi^{(t)}, p_1^{(t)}, p_2^{(t)})$ for the random variable z_i associated with the i^{th} toss? Note that your expression should only involve $\pi^{(t)}, p_1^{(t)}, p_2^{(t)}$ and x_i. For notational convenience, you can drop the superscript t from your expression.

Answer:

$$\begin{aligned} q^*(z) &= p(z|x;\pi^{(t)},p_1^{(t)},p_2^{(t)}) \\ &= \frac{p(z,x;\pi^{(t)},p_1^{(t)},p_2^{(t)})}{p(x;\pi^{(t)},p_1^{(t)},p_2^{(t)})} \\ &= \frac{p(z,x;\pi^{(t)},p_1^{(t)},p_2^{(t)})}{\sum_z p(z,x;\pi^{(t)},p_1^{(t)},p_2^{(t)})} \end{aligned}$$

3. **(4 points)** We can show that the updates for the parameters π, p_1 and p_2 in the M-step are given by

$$\pi^{(t+1)} = \frac{\sum_{i=1}^N E[z_i|x;\pi^{(t)},p_1^{(t)},p_2^{(t)}]}{N} \tag{15}$$

$$p_1^{(t+1)} = \frac{\sum_{i=1}^N E[z_i|x;\pi^{(t)},p_1^{(t)},p_2^{(t)}]x_i}{\sum_{i=1}^N E[z_i|x;\pi^{(t)},p_1^{(t)},p_2^{(t)}]} \tag{16}$$

$$p_2^{(t+1)} = \frac{\sum_{i=1}^N (1-E[z_i|x;\pi^{(t)},p_1^{(t)},p_2^{(t)}])x_i}{\sum_{i=1}^N (1-E[z_i|x;\pi^{(t)},p_1^{(t)},p_2^{(t)}])} \tag{17}$$

How are these updates different from the maximum-likelihood expressions derived for the parameters in the setting where all the variables are observed?

Answer: In the maximum-likelihood expressions, the coin used is known, so the values of z are set. In EM, the updates all used the expected values of z. You can view this as a soft clustering, where in each round, partially of the first coin was used and the other part used the second coin.

4 K-means (programming) [Qirong Ho, 25 points + 5 bonus points]

4.1 K-means finds a local minimum of ψ

1. Define $\{z_1,\ldots,z_n\}$ as the index of closest centers produced in step 2. Let's first see what happens when we change one of those indices. Without loss of generality, assume we change z_p to z'_p where $z'_p \neq z_p$. According to step 2, we also know that

$$||x_p - c_{z'_p}||^2 \geq ||x_p - c_{z_p}||^2$$

Denote ϕ as the objective function corresponding to $\{z_1,\ldots,z_p,\ldots,z_n\}$ and ϕ' as the objective function corresponding to $\{z_1,\ldots,z'_p,\ldots,z_n\}$. Then we have

$$\begin{aligned} & \phi - \phi' \\ = & \sum_{i=1}^n \sum_{j=1}^k \delta(z_i=j)||x_i-c_j||^2 - \sum_{i=1,i\neq p}^n \sum_{j=1}^k \delta(z_i=j)||x_i-c_j||^2 - \sum_{j=1}^k \delta(z'_p=j)||x_p-c_j||^2 \\ = & \sum_{j=1}^k \delta(z_p=j)||x_p-c_j||^2 - \sum_{j=1}^k \delta(z'_p=j)||x_p-c_j||^2 \\ = & ||x_p-c_{z_p}||^2 - ||x_p-c'_{z_p}||^2 \leq 0 \end{aligned}$$

Therefore, the value of ϕ will either increase or stay the same. We could do a similar calculation for simultaneously changing more than one z_i.

2. With z_i fixed, we could optimize $\{c_1,\ldots,c_k\}$ separately. Let's first compute the derivative of ϕ w.r.t.

c_q

$$\begin{aligned}\frac{\partial \phi}{\partial c_q} &= \frac{\partial}{\partial c_q}\left[\sum_{i=1}^{n}\sum_{j=1}^{k}\delta(z_i = j)||x_i - c_j||^2\right] \\ &= \frac{\partial}{\partial c_q}\left[\sum_{i=1}^{n}\delta(z_i = q)||x_i - c_q||^2\right] \\ &= 2\sum_{i=1}^{n}\delta(z_i = q)(c_q - x_i)\end{aligned}$$

Therefore, by setting $\frac{\partial \phi}{\partial c_q} = 0$, we get

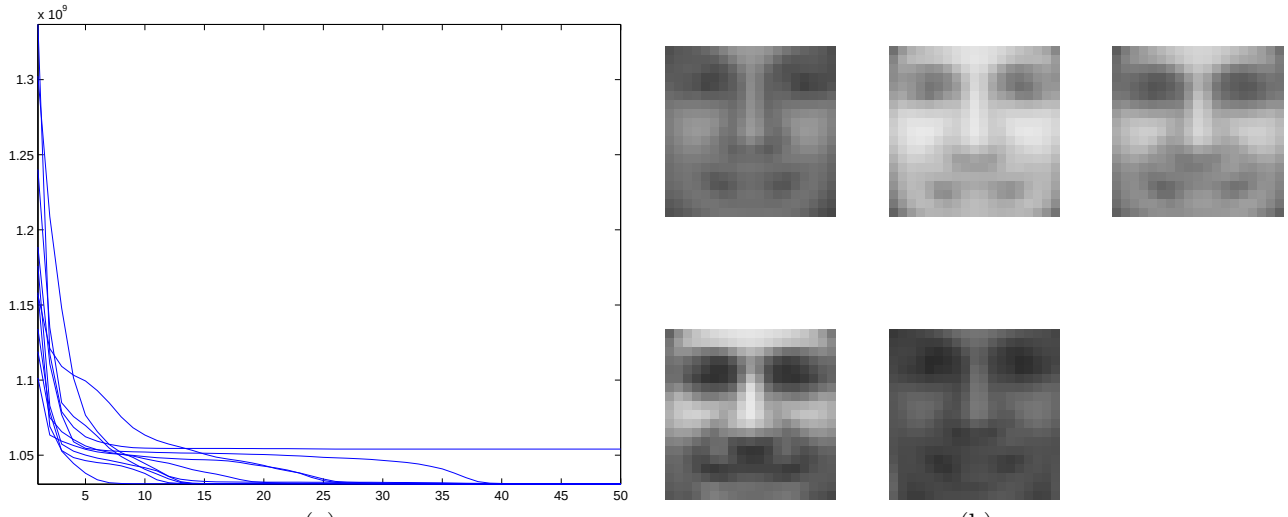
$$c_q = \frac{\sum_{i=1}^{n}\delta(z_i = q)x_i}{\sum_{i=1}^{n}\delta(z_i = q)}$$

Let's further compute the second derivative of ϕ w.r.t. c_q to make sure we are actually minimizing ϕ

$$\frac{\partial^2 \phi}{\partial c_q^2} = \frac{\partial}{\partial c_q}\left[2\sum_{i=1}^{n}\delta(z_i = q)(c_q - x_i)\right] = 2\sum_{i=1}^{n}\delta(z_i = q) \geq 0$$

4.2 Implementing k-means

1. See code.
2. See figure 3(a). Function ϕ is decreasing monotonically, and different runs converge to different values for ϕ. Thus, K-means is only getting local minimum of ϕ.



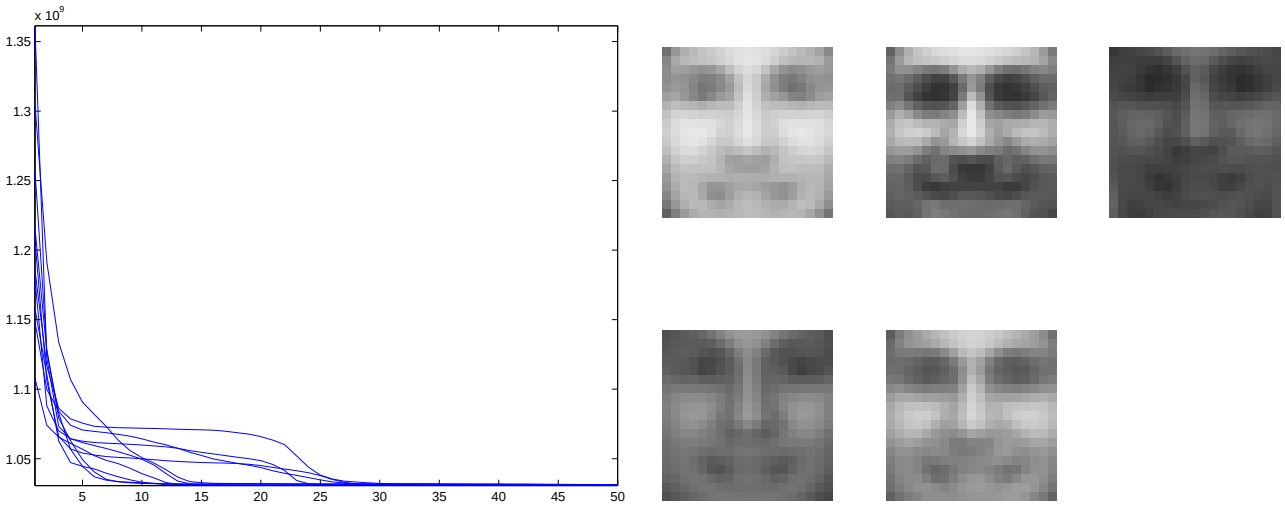
(a) (b)

Figure 3: Plots for face.csv (a) ϕ vs. iteration number (b) visualization of 5 centers.

3. See figure 3(b). Yes, "center-faces" look like faces.

4.3 Bonus Question: choosing good initial centers with the k-means++ algorithm

1. See code.



(a) (b)

Figure 4: Plots for face.csv (a) ϕ vs. iteration number (b) visualization of 5 centers.

2. See figure 4(a).

3. See figure 4(b). Yes, "center-faces" look like faces.

   The main difference is that all runs of k-means++ converge to approximately the same value, which is not true for k-means. K-means++ also converges faster on average.