A

Final Exam
GU4241/GR5241 Fall 2016

Name _____

UNI _____

Problem 0: UNI (2 points)

Write your name and UNI on the first page of the problem sheet. After the exam, please return the problem sheet to us.

Problem 1: Short questions (3+3+4+8 points)

Short answers (about one sentence) are sufficient.

(a) What is the difference between bagging and random forests?

(b) Describe the procedure for computing the out of bag error from Bagging a regression tree.

(c) Assume that f is a convex function. $G = \{\mathbf{x} : g(\mathbf{x}) \leq 0\}$ is a convex set. For the constrained optimization problem:

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g(\mathbf{x}) \leq 0, \end{aligned}$$

the KKT conditions are:

$$\begin{aligned} \nabla f(\mathbf{x}) &= -\lambda \nabla g(\mathbf{x}) \\ \lambda g(\mathbf{x}) &= 0 \\ g(\mathbf{x}) &\leq 0 \\ \lambda &\geq 0 \end{aligned}$$

What is the purpose of each condition? Could you briefly explain the meaning of them?

Solution:

(a) In each case, we average the result of decision trees fit to many Bootstrap samples. However, in a Random Forest we restrict the number of variables to consider in each split. This produces a greater diversity of decision trees or *weak learners*, which reduces the variance of the averaged prediction.

(b) For each sample x_i, consider all the Bootstrap samples that do not contain x_i and average the predictions for the response y_i made by the corresponding trees; call the average $\hat{y}_i^{\text{oob}}$. Then, the out of bag error is given by:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{\text{oob}})^2.$$

(c) The first condition is the optimal criterion. If $\lambda = 0$, this is the criterion for achieving the minimum in the interior of G. If $\lambda > 0$, this means the directions of ∇f and ∇g are opposite, which is the optimal criterion for achieving the minimum at the bounday of G.

The second condition distinguishes the cases when the minimum is achieved in the interior and at the boundary of G.

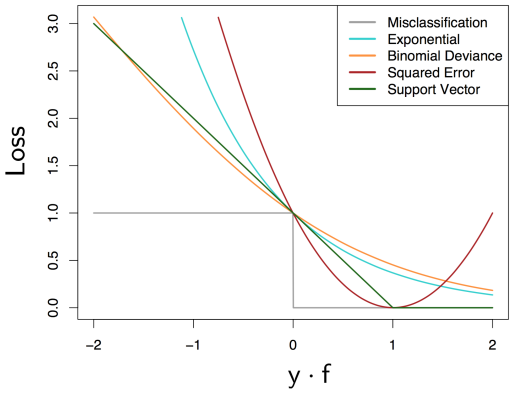
The third condition is the contraint of the problem.

The last one requires that ∇f cannot flip to orientation of ∇g.

(d) Assume that y is a binary response, $y \in \{-1, 1\}$. $f(x)$ is our classifier. The following are the loss functions that are frequently used in machine learning:

- 0-1 Loss: $L_{01}(y, f(x)) = \begin{cases} 0 & yf(x) > 0 \\ 1 & yf(x) \leq 0. \end{cases}$
- Hinge Loss (used in SVM): $L_{\text{h}}(y, f(x)) = \max\{0, 1 - yf(x)\}$.
- Square Loss: $L_{\text{sq}}(y, f(x)) = (1 - yf(x))^2$.
- Exponential Loss (used in boosting): $L_{\text{exp}}(y, f(x)) = \exp(-yf(x))$.
- Binomial Deviance (used in logistic regression): $L_{\text{bi}}(y, f(x)) = \log(1 + \exp(-yf(x)))$.

The plot of these loss functions are shown in the figure below:



- These loss functions can be thought of as convex approximation to the 0-1 loss function. Looking at the plot, which one appears intuitively to be the worst approximation to the 0-1 loss function? Which one appears to be the best?
- Consider just the 0-1 loss, the hinge loss, and the exponential loss. Rank the loss functions from highest to lowest in terms of robustness to misspecification of class labels in the data.
  Hint: consider the amount of loss assigned to a point z that is large and negative; this point corresponds to a large margin error (or a misspecification).

Solution:

(d)
- The squared error loss is the worst approximation, while the hinge loss is the best.
- Exponential loss is the least robust one, and 0-1 loss is the most robust one.

Problem 2: Page Rank (10 points)

Consider a directed graph $G = (V, E)$ with $V = \{1, 2, 3, 4, 5\}$, and $E = \{(1,2), (1,3), (2,1), (2,3), (3,4), (3,5), (4,5), (5,4)\}$ (Note that (a, b) represents an edge from node a to node b). Set up the equations to compute pagerank for G.

Recall in PageRank algorithm, the ranking is given by the invariant distribution of a Markov Chain with transition matrix

$$\mathbf{p} = (1-\alpha)A + \frac{\alpha}{d}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix},$$

where A is the adjacency matrix of the graph defined as

$$A_{ji} = \begin{cases} \frac{1}{\text{degree of node } i} & \text{if } i \text{ links to } j, \\ 0 & \text{otherwise.} \end{cases}$$

d is the number of nodes. If we set $\alpha = 0.2$ here, please write the equations explicitly. DO NOT use matrix representation since we only have five nodes.

Solution:

$$\begin{aligned} r(1) &= 0.4r(2) + 0.04 \\ r(2) &= 0.4r(1) + 0.04 \\ r(3) &= 0.4r(1) + 0.4r(2) + 0.04 \\ r(4) &= 0.4r(3) + 0.8r(5) + 0.04 \\ r(5) &= 0.4r(3) + 0.8r(4) + 0.04 \end{aligned}$$

Problem 3: Support Vector Machines(10 points)

Here we consider an alternative formulation of the soft margin SVM. Given some training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ in $\mathbb{R}^{p+1}$(i.e. each $\mathbf{x}_i$ is a p-dimensional vector and y_i is the one dimensional response), in order to estimate $\mathbf{v}_H$ ($\mathbf{v}_H = (v_H^1, \ldots, v_H^p) \in \mathbb{R}^p$) and b which determine the decision boundary, we can directly minimize the cost:

$$f(\mathbf{v}_H, b) = \frac{1}{2}\|\mathbf{v}_H\|^2 + C\sum_{i=1}^{n} \max\left\{0, 1 - y_i\left(\mathbf{v}_H^T\mathbf{x}_i + b\right)\right\},$$

where $\|\mathbf{v}_H\|^2 = \sum_{j=1}^{p}(v_H^j)^2$. If we define $L(\mathbf{v}_H, b; \mathbf{x}_i, y_i) = \max\left\{0, 1 - y_i\left(\mathbf{v}_H^T\mathbf{x}_i + b\right)\right\}$, then

$$\frac{\partial L}{\partial v_H^j} = \begin{cases} 0 & \text{if } y_i(\mathbf{v}_H^T\mathbf{x}_i + b) \geq 1 \\ -y_i x_i^j & \text{otherwise.} \end{cases}$$

If we want to use gradient descent to solve this optimization problem,

(a) What is $\frac{\partial L(\mathbf{v}_H, b; x_i, y_i)}{\partial b}$?

(b) Write down the gradient update for $\mathbf{v}_H$ and b in each iteration.

Solution:

(a)

$$\frac{\partial L}{\partial b} = \begin{cases} 0 & \text{if } y_i(\mathbf{v}_H^T\mathbf{x}_i + b) \geq 1 \\ -y_i & \text{otherwise.} \end{cases}$$

(b)

$$\begin{aligned} (v_H^j)^{(r+1)} &= (v_H^j)^{(r)} - \left((v_H^j)^{(r)} + C\sum_{i=1}^{n}\frac{\partial L(\mathbf{v}_H, b; x_i, y_i)}{\partial (v_H^j)^{(r)}}\right), \quad \text{for } j = 1, \ldots, p, \\ b^{(r+1)} &= b^{(r)} - \left(C\sum_{i=1}^{n}\frac{\partial L(\mathbf{v}_H, b; x_i, y_i)}{\partial b^{(r)}}\right). \end{aligned}$$

Remark: In practice, we usually need to choose a step size. I.e., the gradient update is as follows:

$$\begin{aligned} (v_H^j)^{(r+1)} &= (v_H^j)^{(r)} - \eta\left((v_H^j)^{(r)} + C\sum_{i=1}^{n}\frac{\partial L(\mathbf{v}_H, b; x_i, y_i)}{\partial (v_H^j)^{(r)}}\right), \quad \text{for } j = 1, \ldots, p, \\ b^{(r+1)} &= b^{(r)} - \eta\left(C\sum_{i=1}^{n}\frac{\partial L(\mathbf{v}_H, b; x_i, y_i)}{\partial b^{(r)}}\right), \end{aligned}$$

where η is the step size. It's between 0 and 1.

Problem 4: Classification Trees(10 points)

We have some data about when people go hiking. The data take into effect, whether hike is on a weekend or not, if the weather is rainy or sunny, and if the person will have company during the hike. Find the optimum decision tree for hiking habits, using the training data below. When you split the decision tree at each node, maximize the drop of impurity by using the entropy as the impurity measure, i.e.,

$$\text{maximize} \quad [I(D) - (I(D_L) + I(D_R))]$$

where D is parent node, and D_L and D_R are two child nodes, and $I(D)$ is:

$$I(D) = mH(\frac{m^+}{m}) = mH(\frac{m^-}{m})$$

and $H(x) = -x\log_2 x - (1-x)\log_2(1-x)$, $0 \leq x \leq 1$, is the entropy function and $m = m^+ + m^-$ with m^+ and m^- being the total number of positive and negative cases at the node.

You may find the following useful in your calculations: $H(x) = H(1-x)$, $H(0) = 0$, $H(1/5) = 0.72$, $H(1/4) = 0.8$, $H(1/3) = 0.92$, $H(2/5) = 0.97$, $H(3/7) = 0.99$, $H(0.5) = 1$.

Weekend?	Company?	Weather	Go Hiking ?
Y	N	R	N
Y	Y	R	N
Y	Y	R	Y
Y	Y	S	Y
Y	N	S	Y
Y	N	S	N
Y	Y	R	N
Y	Y	S	Y
N	Y	S	N
N	Y	R	N
N	N	S	N

(a) Build a decision tree of depth 3. Draw your decision tree.

(b) According to your decision tree, what is the probability of going to hike on a rainy week day, without any company?

(c) How about the probability of going to hike on a rainy weekend when having some company?

Solution:

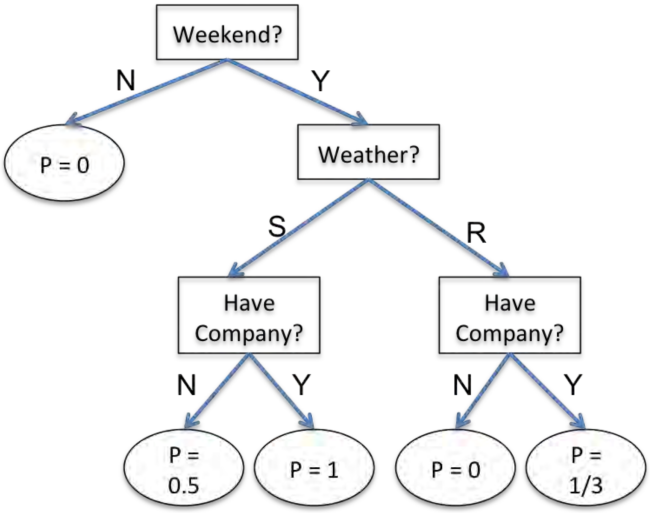
(a) We ant to choose attributes that maximize $mH(p) - m_lH(p_l) - m_r(H(p_r)$, where p, p_l, and p_r are the fraction of positive cases in parent nodes and two child nodes, and m, m_l, and m_r are the total number of points in parent node and two child nodes. This means that at each step, we need to choose the predictor for which $m_rH(p_r) + m_lH(p_l)$ is minimum. For the first step, the predictor *Weekend* achieve this:

- *Weekend*: $m_rH(p_r) + m_lH(p_l) = 8H(1/2) + 3H(0) = 8$
- *Weather*: $m_rH(p_r) + m_lH(p_l) = 5H(1/5) + 6H(1/2) \approx 9.6$
- *Company*: $m_rH(p_r) + m_lH(p_l) = 4H(1/4) + 7H(3/7) \approx 10.1$

Therefore the first split is on *Weekend*. If *Weekend* = N, then the probability of going hiking is 0. If *weekend* = Y, we need to choose the second predictor to split on:

- *Weather*: $m_rH(p_r) + m_lH(p_l) = 4H(1/4) + 4H(1/4) \approx 6.4$
- *Company*: $m_rH(p_r) + m_lH(p_l) = 5H(2/5) + 3H(1/3) \approx 7.6$

Therefore, the second split will be on *Weather*, and the third one will be *Company*. The decision tree is as follows:



(b) Based on the decision tree, the probability is 0.

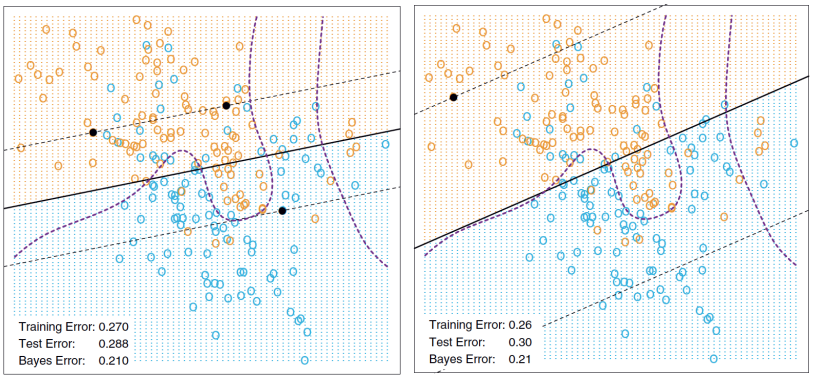
(c) The probability is 1/3.

Problem 5: Decision Boundaries(10 points)

(a) Consider a soft-margin, linear support vector machine:

$$\begin{aligned} \min_{\mathbf{v}_H, b, \xi} \quad & \|\mathbf{v}_H\|^2 + C \sum_{i=1}^{n} \xi_i^2 \\ \text{s.t.} \quad & y_i(\langle \mathbf{v}_H, x_i \rangle - b) \geq 1 - \xi_i, \quad \text{for } i = 1, \ldots, n \\ & \xi_i \geq 0, \quad \text{for } i = 1, \ldots, n. \end{aligned}$$

In the following two plots, the solid lines are the linear decision boundaries obtained by the model above, for two different values of C. The broken lines indicate the margins. Which plot corresponds to a larger value of C? Explain briefly.

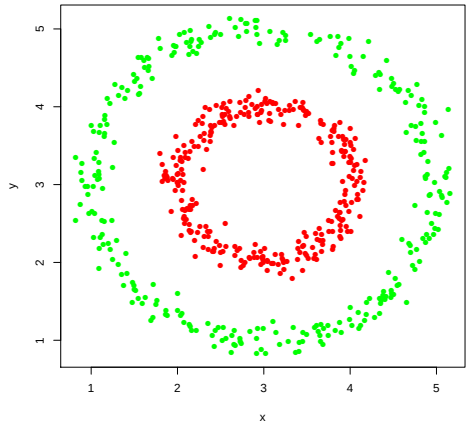


(a) (b)

Solution:

(a) Plot (a) corresponds to a larger value of C. A larger value of C means errors will cost more. Therefore, the margin should be decreased in order to compensate this.

(b) Assume that we have a data set as shown in the following plot, propose at least two classifier which will excel in classifying the data.



Solution:

(b) We can use SVM with a RBF kernel, or logistic regression with x^2, y^2, x and y as features.

Problem 6: Kernelized Nearest Neighbor Classification(10 points)

Since k-nearest-neighbor classifier only computes distances between points, we will exploit this fact to make a kernel version of this classifier, using the same kernel trick that we did in SVM. You may now assume k is given. To derive *kernelized* nearest neighbor classifier, you need follow these steps:

(a) Consider the squared Euclidean distance between any two points $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$. Expand the expression $d^2(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2^2$ in terms of vector inner products.

(b) Assume you are given a kernel $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. Use the kernel trick to make the squared Euclidean distance from the previous part into a kernel squared distance. Write down this expression and call it $d_K^2(\mathbf{x}, \mathbf{x}')$.

(c) d_K is itself a distance measure. What distance does it calculate?

Solution:

(a)

$$d^2(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}', \mathbf{x}' \rangle - 2\langle \mathbf{x}, \mathbf{x}' \rangle$$

(b)

$$\begin{aligned} d_K^2(\mathbf{x}, \mathbf{x}') &= \langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle_{\mathcal{F}} + \langle \phi(\mathbf{x}'), \phi(\mathbf{x}') \rangle_{\mathcal{F}} - 2\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}} \\ &= K(\mathbf{x}, \mathbf{x}) + K(\mathbf{x}', \mathbf{x}') - 2K(\mathbf{x}, \mathbf{x}') \end{aligned}$$

(c) It's the distance between two poins in feature space. $d_K(\mathbf{x}, \mathbf{x}') = \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{F}}$.

Problem 7: Missing Data(10 points)

Assume you are given a data set consisting of variables having more than 30% missing values? Let's say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them? List at least two approaches.

Solution:

(a) We can replace the missing values by the mean of each variable.

(b) We can impute the missing values by running a regression on the other varibles.

(c) We can view this as a matrix completion problem, and use the low rank approximation as our data matrix.

Problem 8: Bayesian Models(10 points)

In this problem, we use the Bayesian approach to estimate the parameters in a simple linear regression. Our model is:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We have a training set $(x_1, y_1), \ldots, (x_n, y_n)$ and want to estimate β_0, β_1, and σ. We will impose a prior on the parameters: the prior on β_0 and β_1 are both Gaussian, and the prior on the INVERSE of σ^2 is Gamma, i.e.,

$$\begin{aligned} \beta_0 &\sim \mathcal{N}(0, \sigma_0^2), \\ \beta_1 &\sim \mathcal{N}(0, \sigma_0^2), \\ \sigma^{-2} &\sim \text{Gamma}(a, b). \end{aligned}$$

Recall that the density of Gaussian distribution is $\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$, and that of Gamma distribution is $g(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$.

Next, we will use the Gibbs sampler to sample from the posterior.

(a) First, we sample from the conditional distribution of β_0. For this parameter, is Gaussian distribution a conjugate prior? Given β_1, σ^2 and the data, does the posterior

$$\Pr\left(\beta_0 | \beta_1, \sigma^2, (x_1, y_1), \ldots, (x_n, y_n)\right)$$

has a closed form expression? If so, what are the parameters of that distribution?

(b) Repeat part (a) for the conditional distribution of σ^{-2}. Does the posterior

$$\Pr\left(\sigma^{-2} | \beta_0, \beta_1, (x_1, y_1), \ldots, (x_n, y_n)\right)$$

has a closed form expression. If so, what are the parameters of that distribution?

Hint:

- Note that the posterior is proportional to prior $\times$ likelihood. Priors are specified as above, and the likelihood of the model is $\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2})$.
- Part (b) might be easier.
- For part (b), you should work with σ^{-2} instead of σ.

Solution:

(a) Denote ϕ as the density of the Gaussian distribution.

$$\begin{aligned} & \Pr\left(\beta_0 | \beta_1, \sigma^2, (x_1, y_1), \ldots, (x_n, y_n)\right) \\ \propto \;& \phi(\beta_0 | 0, \sigma_0^2) \prod_{i=1}^n \phi(y_i - \beta_0 - \beta_1 x_i | 0, \sigma^2) \\ = \;& \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{\beta_0^2}{2\sigma_0^2}\right) \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \\ \propto \;& \exp\left(-\frac{\beta_0^2}{2\sigma_0^2} - \frac{1}{2\sigma^2} \sum_{i=1}^n \left(\beta_0^2 - 2(y_i - \beta_1 x_i)\beta_0\right)\right) \\ = \;& \exp\left(-\left(\frac{1}{2\sigma_0^2} + \frac{n}{2\sigma^2}\right)\beta_0^2 + 2 \cdot \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{2\sigma^2} \beta_0\right). \end{aligned}$$

Therefore, the posterior of β_0 is still Gaussian, with mean

$$\text{posterior mean} = \frac{\frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{2\sigma^2}}{\frac{1}{2\sigma_0^2} + \frac{n}{2\sigma^2}} = \frac{\sigma_0^2(\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i)}{\sigma^2 + n\sigma_0^2},$$

and variance $\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}$.

(b)

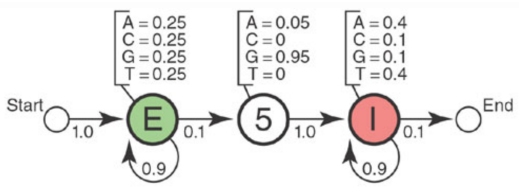
$$\begin{aligned}
& \Pr\left(\sigma^{-2}|\beta_0, \beta_1, (x_1, y_1), \ldots, (x_n, y_n)\right) \\
\propto\ & g(\sigma^{-2}|a, b) \prod_{i=1}^n \phi(y_i - \beta_0 - \beta_1 x_i|0, \sigma^2) \\
=\ & \frac{b^a}{\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a-1} e^{-b/\sigma^2} \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \\
\propto\ & \left(\frac{1}{\sigma^2}\right)^{a+\frac{n}{2}-1} \exp\left(-\left(b + \frac{1}{2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right)/\sigma^2\right)
\end{aligned}$$

Therefore, the posterior of σ^{-2} is still gamma, with shape parameter $a + \frac{n}{2}$ and rate parameter $b + \frac{1}{2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$.

Problem 9: Hidden Markov Models(10 points)

HMMs have several important applications in genetics. In this problem, we consider a 5' splice-site recognition problem in genetics. A DNA sequence is a sequence of amino acids. There are four acids, represented by the symbols A, C, G, and T. The sequence consists of *exon*, *intron*, and the *5' splice site (5'SS)* indicating the switch from exon to intron (assume that there is always a 5'SS between exon and intron). Our goal is to identify where the switch occurred, i.e., where the 5'SS is. We make the following assumptions, which are also shown in the figure below:

- At each symbol in an exon, the probability that the exon ends (i.e., that the next symbol is the 5'SS) is 0.1.
- The 5'SS only contains one symbol, i.e., the probability that the next symbol belongs to an intron is 1.
- At each symbol in an intron, the probability that the intron ends (i.e., that the next symbol belongs to an exon) is 0.1.
- 5'SS can only occur when the sequence switches from exon to intron.
- In an exon, the probabilities of observing A, C, G, or T are all equal to 0.25.
- Introns are A/T rich, say the probability of observing A, C, G or T are 0.4, 0.1, 0.1, and 0.4 respectively.
- At 5'SS, the probability of observing A, C, G, or, T are 0.05, 0, 0.95, and 0 respectively.



Define an HMM for this DNA sequence model. Please specify:

- The state space.
- The set of observations.
- The emission distribution.
- The transition matrix.

Solution:

- State space: $\mathbf{Z} = \{\text{exon}, \text{5'SS}, \text{intron}\}$.
- Observation space: $\mathbf{X} = \{A, C, G, T\}$.
- Emission distributions:

$$\begin{aligned} P(X|Z = \text{exon}) &= \text{Multinomial}(0.25, 0.25, 0.25, 0.25) \\ P(X|Z = \text{intron}) &= \text{Multinomial}(0.4, 0.1, 0.1, 0.4) \\ P(X|Z = \text{5'SS}) &= \text{Multinomial}(0.05, 0, 0.95, 0) \end{aligned}$$

- Transition matrix of the Markov chain:

$$\begin{pmatrix} p_{\text{exon}\to\text{exon}} & p_{\text{5'SS}\to\text{exon}} & p_{\text{intron}\to\text{exon}} \\ p_{\text{exon}\to\text{5'SS}} & p_{\text{5'SS}\to\text{5'SS}} & p_{\text{intron}\to\text{5'SS}} \\ p_{\text{exon}\to\text{intron}} & p_{\text{5'SS}\to\text{intron}} & p_{\text{intron}\to\text{intron}} \end{pmatrix} = \begin{pmatrix} 0.9 & 0 & 0.1 \\ 0.1 & 0 & 0 \\ 0 & 1 & 0.9 \end{pmatrix}$$