CUB Spring 2024. Machine Learning.
Exam test variant.
REFERENCE SOLUTION

Task 1. The right answers here are (c) and (e).

Task 2.

$$df = d\det(\boldsymbol{x}\boldsymbol{y}^T + A) = \{\text{apply formula for differential of determinant}\} =$$

$$= \det(\boldsymbol{x}\boldsymbol{y}^T + A)\mathrm{tr}((\boldsymbol{x}\boldsymbol{y}^T + A)^{-1} d(\boldsymbol{x}\boldsymbol{y}^T + A)) =$$

$$= \det(\boldsymbol{x}\boldsymbol{y}^T + A)\mathrm{tr}((\boldsymbol{x}\boldsymbol{y}^T + A)^{-1} d\boldsymbol{x}\boldsymbol{y}^T) = \{\text{circular property of trace}\} = \det(\boldsymbol{x}\boldsymbol{y}^T + A)\mathrm{tr}(\underbrace{\boldsymbol{y}^T(\boldsymbol{x}\boldsymbol{y}^T + A)^{-1} d\boldsymbol{x}}_{\in\mathbb{R}}) =$$

$$= \underbrace{\det(\boldsymbol{x}\boldsymbol{y}^T + A)\boldsymbol{y}^T(\boldsymbol{x}\boldsymbol{y}^T + A)^{-1}}_{\nabla f(\boldsymbol{x})^T} d\boldsymbol{x}.$$

Hence the answer is

$$\nabla f(\boldsymbol{x}) = \det(\boldsymbol{x}\boldsymbol{y}^T + A)(\boldsymbol{x}\boldsymbol{y}^T + A)^{-T}\boldsymbol{y}.$$

Task 3. The right answer here is only (c).

Task 4. Let's first compute the gradient of loss function for one object:

$$d\log(1 + \exp(-y_i\boldsymbol{w}^T\boldsymbol{x}_i)) = \frac{1}{1 + \exp(-y_i\boldsymbol{w}^T\boldsymbol{x}_i)}\exp(-y_i\boldsymbol{w}^T\boldsymbol{x}_i)d(-y_i\boldsymbol{w}^T\boldsymbol{x}_i) = \frac{1}{1 + \exp(y_i\boldsymbol{w}^T\boldsymbol{x}_i)}(-y_i\boldsymbol{x}_i^T d\boldsymbol{w}).$$

Hence the gradient is equal to

$$\frac{1}{1 + \exp(y_i\boldsymbol{w}^T\boldsymbol{x}_i)}(-y_i\boldsymbol{x}_i).$$

Then one step of SGD is

$$i_k \sim \mathrm{Unif}(1, 2, \ldots, N),$$

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha\frac{1}{1 + \exp(y_{i_k}\boldsymbol{w}_k^T\boldsymbol{x}_{i_k})}(-y_{i_k}\boldsymbol{x}_{i_k}) = \boldsymbol{w}_k + \frac{\alpha y_{i_k}\boldsymbol{x}_{i_k}}{1 + \exp(y_{i_k}\boldsymbol{w}_k^T\boldsymbol{x}_{i_k})}.$$

Task 5. The right answers here are (a) and (b).

Task 6. If classifier makes three errors, then the following configurations are possible:

Below threshold	above threshold	prec	rec	F_1
-1	$[-1, -1, -1, 1, 1, 1]$	1/2	1	2/3
$[-1, -1, 1]$	$[-1, -1, 1, 1]$	1/2	2/3	$< 2/3$
$[-1, -1, -1, 1, 1]$	$[-1, 1]$	1/2	1/3	$< 2/3$
$[-1, -1, -1, -1, 1, 1, 1]$	$\emptyset$	1	0	0

So the maximal possible F_1-measure here is 2/3.

Task 7. Let's sort all objects w.r.t. their prediction score and compute precision, recall and F_1-measure for all thresholds:

score	y_{true}	prec.	rec.	F_1
0.7	1	1	1/3	1/2
0.3	-1	1/2	1/3	2/5
0.1	1	2/3	2/3	2/3
-0.2	1	3/4	1	6/7

If all objects are assigned to negative class then rec = 0 and precision by agreement is supposed to be 1. So F_1-measure for this case is 0. The best F_1 value is 6/7, when all objects are assigned to positive class.

Task 8. Here we need for each feature first sort all the objects. This costs $O(N\log(N))$. Then in a cycle for each threshold we need to reestimate variance of target values for left and right subsets. This costs $O(N)$. So the total computational complexity is $O(DN\log(N))$.

Task 9. We need to solve the following optimization problem:

$$\frac{1}{|R|}\sum_{y\in R}\sum_{k=1}^{K}(c_k-[y=k])^2 \to \min_{c_1,\dots,c_K}.$$

Let's compute a derivative w.r.t. c_j and equate it to zero:

$$\frac{\partial}{\partial c_j}=\frac{1}{|R|}\sum_{y\in R}2(c_j-[y=j])=\frac{2c_j}{|R|}\sum_{y\in R}1-\frac{2}{|R|}\sum_{y\in R}[y=j]=2c_j-\frac{2}{|R|}\sum_{y\in R}[y=j]=0 \;\Rightarrow\; c_{opt,j}=\frac{1}{|R|}\sum_{y\in R}[y=j].$$

Task 10. The right answers here are (a), (c), (d).

Task 11. The right answer here is only (b).

Task 12. The right answers are (a) and (d). Bias term is responsible for complexity of the mean performance algorithm (the difference between mean performance algorithm and the theoretical optimal algorithm). Increasing maximal depth in decision tree increases the complexity of algorithm's family thus reducing the bias. Decreasing regularization coefficient allows the algorithm better fit the training set thus reducing the bias. Changing number of trees in bagging doesn't change bias term because all decision trees have the same complexity.

Task 13. On each iteration of K-means procedure we need to compute distance between each dataset point and each cluster center. Computing Euclidean distance between two vectors of size D costs $O(D)$. In total we need to find $O(NK)$ distances. So here we need in total $O(NKD)$. Then for each dataset point we need to find the closest cluster center. This costs $O(NK)$. Then we need to recompute new cluster centers by averaging points from each cluster. Here we need $O(ND)$.

So the total computational complexity is $O(NKD)$.

Task 14. The right answers are (a) and (d).

Task 15. The only clusterization algorithm in the list is DBSCAN. T-SNE is used for data visualization, LambdaMART is used for ranking and Random Forest is used mostly for regression.

Task 16. Let's write down likelihood function:

$$p(X|\lambda)=\prod_{i=1}^{N}p(x_i|\lambda)=\prod_{i=1}^{N}\exp(-\lambda)\frac{\lambda^{x_i}}{x_i!}.$$

Let's take a logarithm of likelihood function since it doesn't change its optimum:

$$\log p(X|\lambda)=\sum_{i=1}^{N}(-\lambda+x_i\log(\lambda)-\log(x_i!))=-\lambda N+\log(\lambda)\sum_{i=1}^{N}x_i+\text{const}\to\max_{\lambda}.$$

Let's take a derivative w.r.t. λ and equate is to zero:

$$\frac{d}{d\lambda}\log p(X|\lambda)=-N+\frac{\sum_i x_i}{\lambda}=0\;\Rightarrow\;\lambda_{ML}=\frac{\sum_i x_i}{N}.$$

Task 17.

Precision values here: pr@1 $=1$, pr@2 $=\frac{1}{2}$, pr@3 $=\frac{1}{3}$, pr@4 $=\frac{1}{4}$, pr@5 $=\frac{2}{5}$. The corresponding average precision values: AP@1 $=1$, AP@2 $=1$, AP@3 $=1$, AP@4 $=1$, AP@5 $=\frac{1}{2}\cdot 1+\frac{1}{2}\cdot\frac{2}{5}=\frac{7}{10}$.

Task 18. DCG is computed by the following general formula:

$$\text{DCG@k} = \sum_{i=1}^{k} \text{Gain}(y_i)\text{Discount}(i).$$

If $y_2 = 3, y_4 = 4$, then

$$\text{DCG@k} = (2^3 - 1) * \frac{1}{2} + (2^4 - 1) * \frac{1}{4} = \frac{7}{2} + \frac{15}{4} = \frac{29}{4}$$

Normalized DCG computed as follows:

$$\text{nDCG@k} = \frac{DCG@k}{\max DCG@k}.$$

Here $\max DCG@k$ is computed as DCG value for most optimal ranking, i.e. $y_1 = 4, y_2 = 3, y_3 = y_4 = y_5 = 0$:

$$\max \text{DCG@k} = (2^4 - 1) * 1 + (2^3 - 1) * \frac{1}{2} = 15 + \frac{7}{2} = \frac{37}{2}.$$

Finally,

$$\text{nDCG@k} = \frac{DCG@k}{\max DCG@k} = \frac{29}{4} \cdot \frac{2}{37} = \frac{29}{74}$$

Task 19. First of all let's transform optimization problem to the minimization one:

$$\begin{aligned}
&\sum_i p_i \log(p_i) \to \min_{p_1,\dots,p_N},\\
&\sum_i p_i = 1,\\
&p_i \geq 0 \ \forall i.
\end{aligned}$$

Let's write down Lagrange function:

$$L(p_1, \dots, p_N, \mu) = \sum_i p_i \log(p_i) + \mu(\sum_i p_i - 1).$$

Let's find the minimum of Lagrange function w.r.t. p for given μ:

$$\frac{\partial}{\partial p_i} L(\boldsymbol{p}, \mu) = \log(p_i) + 1 + \mu = 0 \ \Rightarrow \ p_{opt,i} = \exp(-1 - \mu).$$

Let's substitute this result into the constraint:

$$1 = \sum_i p_i = \sum_i \exp(-1 - \mu) = N \exp(-1 - \mu) \ \Rightarrow \ \exp(-1 - \mu) = \frac{1}{N} = p_{opt,i}.$$

Hence the distribution with maximal entropy is a uniform distribution.

Task 20. KL divergence between two probability distributions is defined as follows:

$$KL(p||q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}.$$

For discrete distributions integration is changed by summation over all possible values. For given p and q:

$$KL(p||q) = \frac{1}{3} \log(\frac{1}{3} \cdot \frac{6}{1}) + \frac{1}{3} \log(\frac{1}{3} \cdot \frac{3}{1}) + \frac{1}{3} \log(\frac{1}{3} \cdot \frac{2}{1}) = \frac{1}{3} \log(2) + \frac{1}{3} \log \frac{2}{3} = \frac{2}{3} \log(2) - \frac{1}{3} \log(3).$$