ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-21

AUC Maximization in Bayesian Hierarchical Models

Mehmet Gönen[1]

Abstract. The area under the curve (AUC) measures such as the area under the receiver operating characteristics curve (AUROC) and the area under the precision-recall curve (AUPR) are known to be more appropriate than the error rate, especially, for imbalanced data sets. There are several algorithms to optimize AUC measures instead of minimizing the error rate. However, this idea has not been fully exploited in Bayesian hierarchical models owing to the difficulties in inference. Here, we formulate a general Bayesian inference framework, called Bayesian AUC Maximization (BAM), to integrate AUC maximization into Bayesian hierarchical models by borrowing the pairwise and listwise ranking ideas from the information retrieval literature. To showcase our BAM framework, we develop two Bayesian linear classifier variants for two ranking approaches and derive their variational inference procedures. We perform validation experiments on four biomedical data sets to demonstrate the better predictive performance of our framework over its error-minimizing counterpart in terms of average AUROC and AUPR values.

1 INTRODUCTION

In binary classification problems, we are given a sample of N independent and identically distributed training instances $\mathbf{X} = \{\boldsymbol{x}_n \in \mathcal{X}\}_{n=1}^N$ and their class labels $\boldsymbol{y} = \{y_n \in \{-1,+1\}\}_{n=1}^N$. We then use $\mathbf{X}$ and $\boldsymbol{y}$ to learn usually a parametric function that can be used to predict the class labels of unseen test instances. Let $\boldsymbol{f} = \{f_n \in \mathbb{R}\}_{n=1}^N$ be the output values of this parametric function when evaluated on the training instances. The output values can be, for example, posterior probabilities assigned to one of the classes in neural networks or discriminant outputs in support vector machines. During training, the classification parameters used to generate the output values are selected by optimizing an objective function, which usually contains a loss function defined on $\boldsymbol{f}$ and $\boldsymbol{y}$ such as the hinge loss and squared error loss to minimize the expected error rate on test instances.

The error rate is by far the most commonly used performance measure to compare different classification models. However, the area under the curve (AUC) measures such as the area under the receiver operating characteristics curve (AUROC) and the area under the precision-recall curve (AUPR) are better suited to imbalanced binary classification problems. As noted by many earlier studies [5,9,11,21], minimizing the error rate may not lead to better AUC measures. To the best of our knowledge, there is not a full-Bayesian algorithm to optimize AUC measures owing to the difficulties in inference.

In this work, we study AUC maximization for Bayesian hierarchical models and propose a novel inference framework, called Bayesian AUC Maximization (BAM), to optimize AUC measures with a full-Bayesian treatment. To this aim, we borrow the pairwise and listwise ranking ideas from the information retrieval literature and show how they can help us maximize AUROC values in Bayesian hierarchical models. We demonstrate the better predictive performance of our framework on four biomedical data sets by comparing it to an error-minimizing baseline algorithm.

[1] Department of Industrial Engineering, Koç University, İstanbul, Turkey, email: mehmetgonen@ku.edu.tr

2 MODELING AUC MAXIMIZATION USING CATEGORICAL DISTRIBUTIONS

To be able to model AUC maximization in Bayesian hierarchical models, we first write AUROC as a function of the output values and then show two possible strategies to represent this function using random variables from the categorical distributions (also known as generalized Bernoulli distribution or multinomial distribution with a single trial).

It is very well-known that AUROC is equal to the value of the Wilcoxon-Mann-Whitney statistic in the discrete case [7]:

$$\text{AUROC}(\boldsymbol{f}) = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{n \in \mathcal{P}} \sum_{o \in \mathcal{N}} \delta(f_n > f_o),$$

where $\mathcal{P} = \{n \colon y_n = +1\}$, $\mathcal{N} = \{n \colon y_n = -1\}$, $|\cdot|$ gives the cardinality of the input set, and $\delta(\cdot)$ represents the Kronecker delta function that returns 1 if its argument is true and 0 otherwise.

The first strategy to represent AUROC using the categorical distributions is similar to the pairwise ranking models in the information retrieval literature, which force the output value of a relevant document to be larger than that of an irrelevant document for all relevant–irrelevant document pairs [3, 6, 10]. The Wilcoxon-Mann-Whitney statistic considers all pairs defined between the positive and negative instances, which we can represent using auxiliary random variables drawn from categorical distributions with two possible outcomes and their respective probabilities calculated using the softmax function:

$$z_{n,o} | f_n, f_o \sim \mathcal{C}\left(z_{n,o}; \{n, o\}, \frac{[\exp(f_n) \quad \exp(f_o)]}{\exp(f_n) + \exp(f_o)} \right), \quad (1)$$

where $(n, o) \in \mathcal{P} \times \mathcal{N}$ and $\mathcal{C}(\cdot; \mathbf{E}, \boldsymbol{\pi})$ denotes the categorical distribution with the event set $\mathbf{E}$ and the event probabilities $\boldsymbol{\pi}$. Figure 1 illustrates the pairwise ranking idea applied on a toy data set with two positive and two negative instances, which requires four categorical random variables to be defined. To maximize AUROC during training, we can treat $z_{n,o}$ variables as observed variables and set each of them to the index of the positive instance involved, i.e., $z_{n,o} = n$.

The pairwise ranking strategy requires $|\mathcal{P}| \times |\mathcal{N}|$ categorical random variables to be added, whereas we can reduce this number to $\min(|\mathcal{P}|, |\mathcal{N}|)$ using the listwise ranking models in the information retrieval literature, which force the output value of a relevant document to be larger than those of all irrelevant documents at once

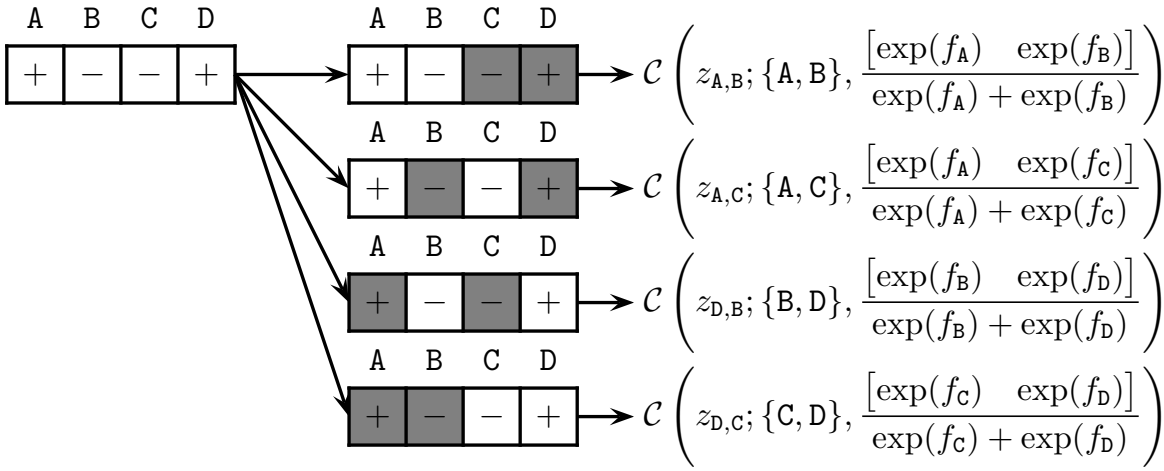

Figure 1. Pairwise ranking applied to modeling AUC maximization using categorical distributions.

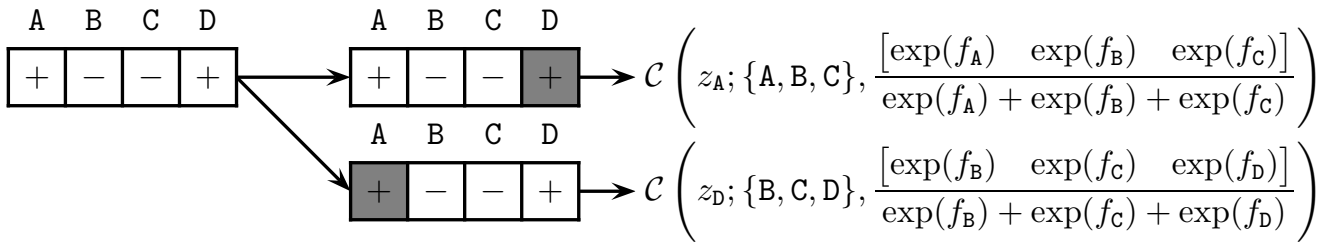

Figure 2. Listwise ranking applied to modeling AUC maximization using categorical distributions.

[4, 13, 20]. Without loss of generality, we assume that $|\mathcal{P}| < |\mathcal{N}|$ in the following. For each positive instance, we can add an auxiliary random variable drawn from a categorical distribution with $(|\mathcal{N}|+1)$ possible outcomes and their respective probabilities calculated using the softmax function:

$$z_n|\{f_o\}_{o\in\mathcal{W}_n} \sim \mathcal{C}\left(z_n;\mathcal{W}_n,\left[\frac{\exp(f_o)}{\sum_{p\in\mathcal{W}_n}\exp(f_p)}\right]_{o\in\mathcal{W}_n}\right), \quad (2)$$

where $n\in\mathcal{P}$ and $\mathcal{W}_n=\{n\}\cup\mathcal{N}$. Figure 2 illustrates the listwise ranking idea applied on a toy data set with two positive and two negative instances, which requires two categorical random variables to be defined. To maximize AUROC during training, we can treat z_n variables as observed variables and set each of them to the index of the positive instance involved, i.e., $z_n=n$. Reducing the number of categorical random variables from $|\mathcal{P}|\times|\mathcal{N}|$ to $\min(|\mathcal{P}|,|\mathcal{N}|)$ would help us make our inference procedures more effective as detailed later.

3 BAYESIAN AUC MAXIMIZATION

To showcase our framework, without loss of generality, we use Bayesian probit regression model as our baseline method [1]. We first describe this model briefly and then give detailed derivations for our two AUC-maximizing variants of this model.

3.1 Bayesian Probit Regression as Baseline Linear Classifier

The distributional assumptions of Bayesian probit regression model are defined as

$$\begin{aligned}
\gamma &\sim \mathcal{G}(\gamma;\alpha_\gamma,\beta_\gamma),\\
b|\gamma &\sim \mathcal{N}(b;0,\gamma^{-1}),\\
\eta_d &\sim \mathcal{G}(\eta_d;\alpha_\eta,\beta_\eta) && \forall d,\\
w_d|\eta_d &\sim \mathcal{N}(w_d;0,\eta_d^{-1}) && \forall d,\\
f_n|b,\boldsymbol{w},\boldsymbol{x}_n &\sim \mathcal{N}(f_n;\boldsymbol{w}^\top\boldsymbol{x}_n+b,1) && \forall n,\\
y_n|f_n &\sim \delta(f_n y_n>\nu) && \forall n, \quad (3)
\end{aligned}$$

where $\{f_n\}_{n=1}^N$ is the set of output values introduced to make the inference procedures efficient [1]. The nonnegative margin parameter ν is introduced to resolve the scaling ambiguity and to place a low-density region between two classes, similar to the margin idea in support vector machines. $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ represents the normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. $\mathcal{G}(\cdot;\alpha,\beta)$ denotes the gamma distribution with the shape parameter α and the scale parameter β.

We can approximate the required posterior as

$$p(\gamma,\boldsymbol{\eta},b,\boldsymbol{w},\boldsymbol{f}|\mathbf{X},\boldsymbol{y}) \approx q(\gamma,\boldsymbol{\eta},b,\boldsymbol{w},\boldsymbol{f}) = q(\gamma)q(\boldsymbol{\eta})q(b,\boldsymbol{w})q(\boldsymbol{f}),$$

and define each factor in the ensemble just like its full conditional distribution:

$$\begin{aligned}
q(\gamma) &= \mathcal{G}(\gamma;\alpha(\gamma),\beta(\gamma)),\\
q(\boldsymbol{\eta}) &= \prod_{d=1}^{D}\mathcal{G}(\eta_d;\alpha(\eta_d),\beta(\eta_d)),\\
q(b,\boldsymbol{w}) &= \mathcal{N}\left(\begin{bmatrix}b\\\boldsymbol{w}\end{bmatrix};\mu(b,\boldsymbol{w}),\Sigma(b,\boldsymbol{w})\right),\\
q(\boldsymbol{f}) &= \prod_{n=1}^{N}\mathcal{TN}(f_n;\mu(f_n),\Sigma(f_n),\rho(f_n)),
\end{aligned}$$

where $\alpha(\cdot)$, $\beta(\cdot)$, $\mu(\cdot)$, and $\Sigma(\cdot)$ denote the shape parameter, the scale parameter, the mean vector, and the covariance matrix for their arguments, respectively. $\mathcal{TN}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\rho(\cdot))$ represents the truncated normal distribution with the mean vector $\boldsymbol{\mu}$, the covariance matrix $\boldsymbol{\Sigma}$, and the truncation rule $\rho(\cdot)$ such that $\mathcal{TN}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\rho(\cdot)) \propto \mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ if $\rho(\cdot)$ is true and $\mathcal{TN}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\rho(\cdot)) = 0$ otherwise.

We can bound the likelihood using Jensen's inequality:

$$\begin{aligned}\log p(\boldsymbol{y}|\mathbf{X}) &\geq \mathbb{E}_q[\log p(\gamma,\boldsymbol{\eta},b,\boldsymbol{w},\boldsymbol{f},\boldsymbol{y}|\mathbf{X})] - \mathbb{E}_q[\log q(\gamma,\boldsymbol{\eta},b,\boldsymbol{w},\boldsymbol{f})],\end{aligned}$$

where $\mathbb{E}_q[\cdot]$ denotes the posterior expectations, and optimize this bound by maximizing with respect to each factor until convergence, leading to the following update equations:

$$\alpha(\gamma) = \alpha_\gamma + \frac{1}{2},\ \beta(\gamma) = \left(\beta_\gamma^{-1} + \frac{1}{2}\,\mathbb{E}_q[b^2]\right)^{-1}, \tag{4}$$

$$\alpha(\eta_d) = \alpha_\eta + \frac{1}{2},\ \beta(\eta_d) = \left(\beta_\eta^{-1} + \frac{1}{2}\,\mathbb{E}_q[w_d^2]\right)^{-1}, \tag{5}$$

$$\Sigma(b,\boldsymbol{w}) = \begin{bmatrix}\mathbb{E}_q[\gamma] + N & \mathbf{1}^\top\mathbf{X}^\top \\ \mathbf{X}\mathbf{1} & \operatorname{diag}(\mathbb{E}_q[\boldsymbol{\lambda}]) + \mathbf{X}\mathbf{X}^\top\end{bmatrix}^{-1}, \tag{6}$$

$$\mu(b,\boldsymbol{w}) = \Sigma(b,\boldsymbol{w})\left(\begin{bmatrix}\mathbf{1}^\top \\ \mathbf{X}\end{bmatrix}\mathbb{E}_q[\boldsymbol{f}]\right), \tag{7}$$

$$\Sigma(f_n) = 1, \quad \rho(f_n) \triangleq f_n y_n > \nu, \tag{8}$$

$$\mu(f_n) = \Sigma(f_n)\,\mathbb{E}_q[\boldsymbol{w}^\top\boldsymbol{x}_n + b], \tag{9}$$

where $\mathbf{1}$ denotes a vector of ones of appropriate dimension.

3.2 AUC-Maximizing Linear Classifier Variant Using Pairwise Ranking

To develop our linear classification variant with pairwise ranking flavor, we first start by replacing (3) in our baseline Bayesian probit regression model with (1). In this modified model, the update equations for γ, $\boldsymbol{\eta}$, b, and $\boldsymbol{w}$ remain intact because they are assumed to be independent from $\boldsymbol{f}$ in the posterior approximation. However, we can not have closed-form update equations for the parameters of the output values $\{f_n\}_{n=1}^N$ owing to the softmax function in (1). Instead, we update these parameters by solving a series of unconstrained optimization problems on the lower bound.

The lower bound we need to maximize to update the parameters of the output values is

$$\begin{aligned}\mathcal{L}_q(\boldsymbol{f}) = \sum_{n=1}^N &(\mathbb{E}_q[\log p(f_n|b,\boldsymbol{w},\boldsymbol{x}_n)] - \mathbb{E}_q[\log q(f_n)]) \\ &+ \sum_{n\in\mathcal{P}}\sum_{o\in\mathcal{N}} \mathbb{E}_q[\log p(z_{n,o}|f_n,f_o)] + \text{const.},\end{aligned}$$

where the first two terms are log-likelihood and negative entropy terms for the output values, whereas the third term stems from the auxiliary random variables introduced in (1). The log-likelihood term can be decomposed as

$$\begin{aligned}&\mathbb{E}_q[\log p(f_n|b,\boldsymbol{w},\boldsymbol{x}_n)] \\ &= \mathbb{E}_q\left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}(f_n - \boldsymbol{w}^\top\boldsymbol{x}_n - b)^2\right] \\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\,\mathbb{E}_q[f_n^2] + \mathbb{E}_q[f_n]\,\mathbb{E}_q[\boldsymbol{w}^\top\boldsymbol{x}_n + b] \\ &\quad - \frac{1}{2}\,\mathbb{E}_q[(\boldsymbol{w}^\top\boldsymbol{x}_n + b)^2],\end{aligned}$$

where $\mathbb{E}_q[f_n] = \mu(f_n)$ and $\mathbb{E}_q[f_n^2] = \mu(f_n)^2 + \Sigma(f_n)$. The negative entropy term is

$$\mathbb{E}_q[\log q(f_n)] = -\frac{1}{2}(\log(2\pi\Sigma(f_n)) + 1).$$

The last term with the softmax function can be lower bounded with a local variational approximation, which uses a linear Taylor expansion of the log function [2]:

$$\begin{aligned}\mathbb{E}_q[\log p(z_{n,o}|f_n,f_o)] &= \mathbb{E}_q\left[\log\left(\frac{\exp(f_n)}{\exp(f_n)+\exp(f_o)}\right)\right] \\ &= \mathbb{E}_q[f_n] - \mathbb{E}_q[\log(\exp(f_n)+\exp(f_o))] \\ &\geq \mathbb{E}_q[f_n] - \left(\frac{(\mathbb{E}_q[\exp(f_n)] + \mathbb{E}_q[\exp(f_o)])}{\zeta_{n,o}} - 1 + \log(\zeta_{n,o})\right),\end{aligned}$$

where $\{\zeta_{n,o}\}_{n\in\mathcal{P},o\in\mathcal{N}}$ is the set of variational parameters introduced and $\mathbb{E}_q[\exp(f_n)] = \exp(\mu(f_n) + \Sigma(f_n)/2)$.

We can now write the lower bound as a function of $\{\mu(f_n)\}_{n=1}^N$, $\{\Sigma(f_n)\}_{n=1}^N$, and $\{\zeta_{n,o}\}_{n\in\mathcal{P},o\in\mathcal{N}}$, and optimize the lower bound with respect to each of these three sets of parameters separately.

3.2.1 Optimizing $\mathcal{L}_q(\boldsymbol{f})$ with respect to $\{\zeta_{n,o}\}_{n\in\mathcal{P},o\in\mathcal{N}}$

Given $\{\mu(f_n)\}_{n=1}^N$ and $\{\Sigma(f_n)\}_{n=1}^N$, the optimal values for $\{\zeta_{n,o}\}_{n\in\mathcal{P},o\in\mathcal{N}}$ can be found as

$$\zeta_{n,o}^\star = \exp\left(\mu(f_n) + \frac{\Sigma(f_n)}{2}\right) + \exp\left(\mu(f_o) + \frac{\Sigma(f_o)}{2}\right). \tag{10}$$

3.2.2 Derivatives of $\mathcal{L}_q(\boldsymbol{f})$ with respect to $\{\mu(f_n)\}_{n=1}^N$

Given $\{\Sigma(f_n)\}_{n=1}^N$ and $\{\zeta_{n,o}\}_{n\in\mathcal{P},o\in\mathcal{N}}$, we can find the first and second derivatives of the lower bound with respect to $\{\mu(f_n)\}_{n\in\mathcal{P}}$ as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}_q(\boldsymbol{f})}{\partial\mu(f_n)} &= -\mu(f_n) + \mathbb{E}_q[\boldsymbol{w}^\top\boldsymbol{x}_n + b] + |\mathcal{N}| \\ &\quad - \sum_{o\in\mathcal{N}}\frac{1}{\zeta_{n,o}}\exp\left(\mu(f_n) + \frac{\Sigma(f_n)}{2}\right) \\ \frac{\partial^2\mathcal{L}_q(\boldsymbol{f})}{\partial\mu(f_n)^2} &= -1 - \sum_{o\in\mathcal{N}}\frac{1}{\zeta_{n,o}}\exp\left(\mu(f_n) + \frac{\Sigma(f_n)}{2}\right),\end{aligned}$$

and with respect to $\{\mu(f_o)\}_{o\in\mathcal{N}}$ as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}_q(\boldsymbol{f})}{\partial\mu(f_o)} &= -\mu(f_o) + \mathbb{E}_q[\boldsymbol{w}^\top\boldsymbol{x}_o + b] \\ &\quad - \sum_{n\in\mathcal{P}}\frac{1}{\zeta_{n,o}}\exp\left(\mu(f_o) + \frac{\Sigma(f_o)}{2}\right), \\ \frac{\partial^2\mathcal{L}_q(\boldsymbol{f})}{\partial\mu(f_o)^2} &= -1 - \sum_{n\in\mathcal{P}}\frac{1}{\zeta_{n,o}}\exp\left(\mu(f_o) + \frac{\Sigma(f_o)}{2}\right).\end{aligned}$$

3.2.3 Derivatives of $\mathcal{L}_q(\boldsymbol{f})$ with respect to $\{\Sigma(f_n)\}_{n=1}^N$

Given $\{\mu(f_n)\}_{n=1}^N$ and $\{\zeta_{n,o}\}_{n\in\mathcal{P},o\in\mathcal{N}}$, we can find the first and second derivatives of the lower bound with respect to $\{\Sigma(f_n)\}_{n\in\mathcal{P}}$ as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}_q(\boldsymbol{f})}{\partial\Sigma(f_n)} &= -\frac{1}{2} - \sum_{o\in\mathcal{N}}\frac{1}{2\zeta_{n,o}}\exp\left(\mu(f_n) + \frac{\Sigma(f_n)}{2}\right) + \frac{1}{2\Sigma(f_n)}, \\ \frac{\partial^2\mathcal{L}_q(\boldsymbol{f})}{\partial\Sigma(f_n)^2} &= -\sum_{o\in\mathcal{N}}\frac{1}{4\zeta_{n,o}}\exp\left(\mu(f_n) + \frac{\Sigma(f_n)}{2}\right) - \frac{1}{2\Sigma(f_n)^2},\end{aligned}$$

and with respect to $\{\mu(f_o)\}_{o\in\mathcal{N}}$ as follows:

$$\frac{\partial \mathcal{L}_q(\boldsymbol{f})}{\partial \Sigma(f_o)} = -\frac{1}{2} - \sum_{n\in\mathcal{P}} \frac{1}{2\zeta_{n,o}} \exp\left(\mu(f_o) + \frac{\Sigma(f_o)}{2}\right) + \frac{1}{2\Sigma(f_o)},$$

$$\frac{\partial^2 \mathcal{L}_q(\boldsymbol{f})}{\partial \Sigma(f_o)^2} = -\sum_{n\in\mathcal{P}} \frac{1}{4\zeta_{n,o}} \exp\left(\mu(f_o) + \frac{\Sigma(f_o)}{2}\right) - \frac{1}{2\Sigma(f_o)^2}.$$

As our overall inference scheme, we first perform closed-form variational updates for γ, $\boldsymbol{\eta}$, b, and $\boldsymbol{w}$ as given in (4)–(7) at each iteration. However, to optimize the parameters of the output values $\boldsymbol{f}$, we replace (8)–(9) with a series of unconstrained optimization problems. At each iteration, we perform the following four steps to update these parameters:

(i) update $\{\zeta_{n,o}\}_{n\in\mathcal{P},o\in\mathcal{N}}$ using (10),

(ii) optimize $\{\mu(f_n)\}_{n=1}^N$ using $\{\partial\mathcal{L}_q(\boldsymbol{f})/\partial\mu(f_n)\}_{n=1}^N$ and $\{\partial^2\mathcal{L}_q(\boldsymbol{f})/\partial\mu(f_n)^2\}_{n=1}^N$,

(iii) update $\{\zeta_{n,o}\}_{n\in\mathcal{P},o\in\mathcal{N}}$ using (10),

(iv) optimize $\{\Sigma(f_n)\}_{n=1}^N$ using $\{\partial\mathcal{L}_q(\boldsymbol{f})/\partial\Sigma(f_n)\}_{n=1}^N$ and $\{\partial^2\mathcal{L}_q(\boldsymbol{f})/\partial\Sigma(f_n)^2\}_{n=1}^N$.

In the steps (ii) and (iv), we use minFunc Matlab package by Mark Schmidt, which uses a quasi-Newton strategy with limited-memory BFGS updates and is publicly available at https://goo.gl/Vrd5DL. Note that the step (iv) needs to be performed in the log-domain to ensure the non-negativity of the variance parameters.

3.3 AUC-Maximizing Linear Classifier Variant Using Listwise Ranking

We follow the same strategy to develop our linear classification variant with listwise ranking flavor and start by replacing (3) in our baseline Bayesian probit regression model with (2). Different from the pairwise ranking variant, we now need fewer auxiliary random variables, and the lower bound becomes

$$\mathcal{L}_q(\boldsymbol{f}) = \sum_{n=1}^N (\mathbb{E}_q[\log p(f_n|b,\boldsymbol{w},\boldsymbol{x}_n)] - \mathbb{E}_q[\log q(f_n)]) + \sum_{n\in\mathcal{P}} \mathbb{E}_q[\log p(z_n|\{f_o\}_{o\in\mathcal{W}_n})] + \text{const.},$$

where the first two terms are the same as before. The third term with the softmax function can again be lower bounded with a local variational approximation:

$$\begin{aligned}\mathbb{E}_q[\log p(z_n|\{f_o\}_{o\in\mathcal{W}_n})] &= \mathbb{E}_q\left[\log\left(\frac{\exp(f_n)}{\sum_{o\in\mathcal{W}_n}\exp(f_o)}\right)\right] \\ &= \mathbb{E}_q[f_n] - \mathbb{E}_q\left[\log\left(\sum_{o\in\mathcal{W}_n}\exp(f_o)\right)\right] \\ &\geq \mathbb{E}_q[f_n] - \left(\sum_{o\in\mathcal{W}_n}\frac{1}{\zeta_n}\mathbb{E}_q[\exp(f_o)] - 1 + \log(\zeta_n)\right),\end{aligned}$$

where $\{\zeta_n\}_{n\in\mathcal{P}}$ is the set of variational parameters introduced. We expect the lower bound approximation in the listwise setting with significantly fewer variational parameters to be much tighter than that of the pairwise setting.

We can again write the lower bound as a function of the mean, covariance, and variational parameters, and optimize them separately.

3.3.1 *Optimizing $\mathcal{L}_q(\boldsymbol{f})$ with respect to $\{\zeta_n\}_{n\in\mathcal{P}}$*

Given $\{\mu(f_n)\}_{n=1}^N$ and $\{\Sigma(f_n)\}_{n=1}^N$, the optimal values for $\{\zeta_n\}_{n\in\mathcal{P}}$ can be found as

$$\zeta_n^\star = \sum_{o\in\mathcal{W}_n} \exp\left(\mu(f_o) + \frac{\Sigma(f_o)}{2}\right). \tag{11}$$

3.3.2 *Derivatives of $\mathcal{L}_q(\boldsymbol{f})$ with respect to $\{\mu(f_n)\}_{n=1}^N$*

Given $\{\Sigma(f_n)\}_{n=1}^N$ and $\{\zeta_n\}_{n\in\mathcal{P}}$, we can find the first and second derivatives of the lower bound with respect to $\{\mu(f_n)\}_{n=1}^N$ as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}_q(\boldsymbol{f})}{\partial \mu(f_n)} &= -\mu(f_n) + \mathbb{E}_q[\boldsymbol{w}^\top\boldsymbol{x}_n + b] + \delta(y_n = +1) \\ &\quad - \sum_{o\in\mathcal{B}_n}\frac{1}{\zeta_o}\exp\left(\mu(f_n) + \frac{\Sigma(f_n)}{2}\right), \\ \frac{\partial^2 \mathcal{L}_q(\boldsymbol{f})}{\partial \mu(f_n)^2} &= -1 - \sum_{o\in\mathcal{B}_n}\frac{1}{\zeta_o}\exp\left(\mu(f_n) + \frac{\Sigma(f_n)}{2}\right),\end{aligned}$$

where $\mathcal{B}_n$ is $\mathcal{P}$ if $y_n = -1$ and $\{n\}$ otherwise.

3.3.3 *Derivatives of $\mathcal{L}_q(\boldsymbol{f})$ with respect to $\{\Sigma(f_n)\}_{n=1}^N$*

Given $\{\mu(f_n)\}_{n=1}^N$ and $\{\zeta_n\}_{n\in\mathcal{P}}$, we can find the first and second derivatives of the lower bound with respect to $\{\Sigma(f_n)\}_{n=1}^N$ as follows:

$$\frac{\partial \mathcal{L}_q(\boldsymbol{f})}{\partial \Sigma(f_n)} = -\frac{1}{2} - \sum_{o\in\mathcal{B}_n}\frac{1}{2\zeta_o}\exp\left(\mu(f_n) + \frac{\Sigma(f_n)}{2}\right) + \frac{1}{2\Sigma(f_n)},$$

$$\frac{\partial^2 \mathcal{L}_q(\boldsymbol{f})}{\partial \Sigma(f_n)^2} = -\sum_{o\in\mathcal{B}_n}\frac{1}{4\zeta_o}\exp\left(\mu(f_n) + \frac{\Sigma(f_n)}{2}\right) - \frac{1}{2\Sigma(f_n)^2}.$$

Different from our pairwise variant, we now need to update $\{\zeta_n\}_{n\in\mathcal{P}}$ using (11) in the steps (i) and (iii) during the variational inference.

4 EXPERIMENTS

To illustrate the effectiveness of our AUC-maximizing variants with pairwise ranking (BAM_P) and with listwise ranking (BAM_L), we report their results on four biomedical data sets (i.e., two cancer and two HIV data sets) and compare them to the error-minimizing baseline algorithm (i.e., Bayesian probit regression; BPROBIT). We implement these three algorithms in Matlab, and our implementations are publicly available at https://goo.gl/DYh7ZR.

We use AUROC and AUPR values from repeated random subsampling validation experiments to compare the classification performance of the algorithms. For each data set, we create 100 random train/test splits to obtain robust results. For each replication, the training set is defined by randomly selecting 75% of the data points with stratification on the phenotype, and the remaining 25% of the samples are used as the test set. The training set is normalized to have

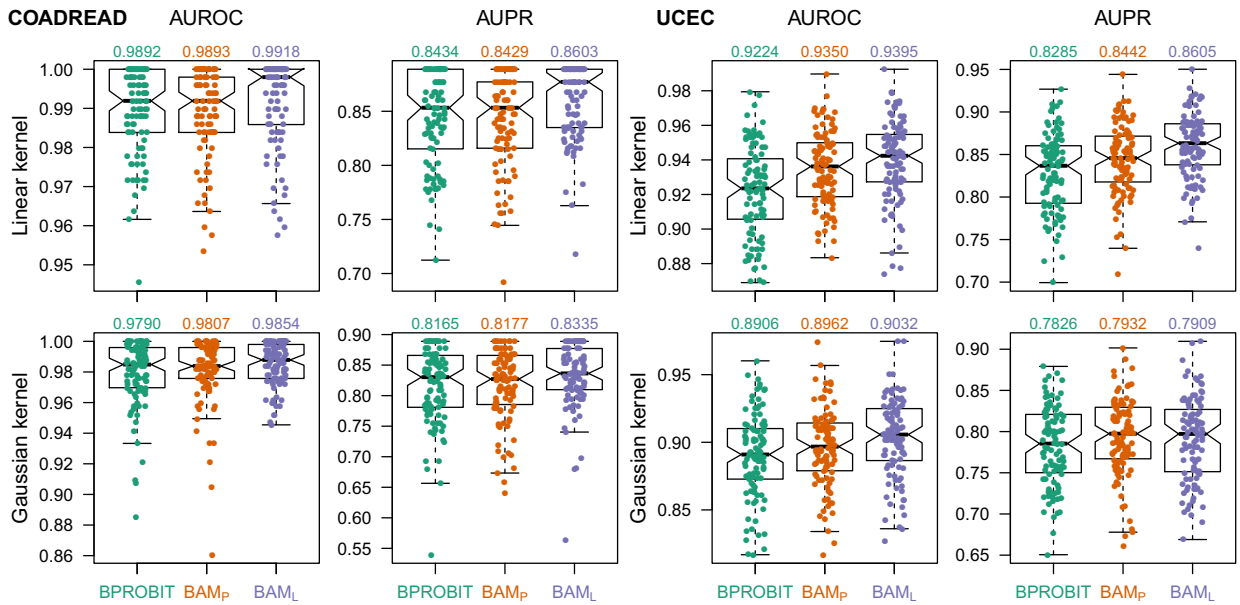

Figure 3. Classification results on two cancer data sets. The box-and-whisker plots show the results of the error-minimizing baseline algorithm (BPROBIT), our AUC-maximizing variant with pairwise ranking (BAM_P), and our AUC-maximizing variant with listwise ranking (BAM_L) over 100 replications in repeated random subsampling validation experiments. The numbers above the figures give the average performance values for each experiment.

zero mean and unit standard deviation, and the test set is then normalized using the mean and the standard deviation of the original training set.

Owing to the high dimensional inputs in our applications, we represent data points using an empirical kernel map, i.e., replacing $\boldsymbol{x}_n$ with $\begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_n) & \dots & k(\boldsymbol{x}_N, \boldsymbol{x}_n) \end{bmatrix}^\top$, which is the main idea behind relevance vector machines [18]. This step reduces the dimensionality of the input space from D to N. We perform experiments with the linear and Gaussian kernels, where we normalize the linear kernel to have unit diagonal entries, and the kernel width of the Gaussian kernel is selected as the average pairwise distance between the training instances.

The hyper-parameter values are selected as $(\alpha_\gamma, \beta_\gamma) = (1, 1)$ and $(\alpha_\eta, \beta_\eta) = (1, 1)$ for all algorithms, and $\nu = 1$ for BPROBIT. We perform at most 200 iterations or stop when the improvement in the lower bound between successive iterations is less than 0.001% during variational inference.

4.1 Classification Results on Cancer Data Sets

Micro-satellite instability is a hypermutable phenotype caused by the loss of DNA mismatch repair activity. It is frequently observed in several tumor types such as colorectal, endometrial, gastric, ovarian, and sebaceous carcinomas [19]. Tumors with micro-satellite instability do not respond to chemotherapeutic strategies developed for micro-satellite stable tumors, leading to its clinical importance. That is why we address the problem of predicting micro-satellite instability status of cancer patients from their gene expression data. We use two publicly available data sets provided by the Cancer Genome Atlas (TCGA) consortium: (i) colon and rectum adenocarcinoma (COADREAD) patients [16] and (ii) uterine corpus endometrial carcinoma (UCEC) patients [17].

The phenotype values of cancer patients for both data sets are downloaded from the TCGA website (https://tcga-data.nci.nih.gov), which groups the patients into three categories: (i) micro-satellite instability high (MSI-H), (ii) micro-satellite instability low (MSI-L), and (iii) micro-satellite stable (MSS). The preprocessed genomic characterizations of primary tumors from the patients (i.e., mRNA gene expression) are downloaded from http://dx.doi.org/10.7303/syn300013, where 20530 normalized gene expression intensities are provided for each profiled primary tumor. We remove the patients with missing phenotype value or genomic data from further analysis. At the end, there are 261 (37 MSI-H, 43 MSI-L, and 181 MSS) and 330 (108 MSI-H, 27 MSI-L, and 195 MSS) patients with available phenotype value and genomic data for COADREAD and UCEC data sets, respectively. We run binary classification experiments to separate MSI-H patients from others (i.e., MSI-L and MSS), which is in agreement with the earlier studies that combine MSI-L and MSS tumors into the same group [19].

Figure 3 compares the performance of the baseline algorithm and our two variants on two cancer data sets in terms of AUROC and AUPR over 100 replications, and reports the average AUROC and AUPR values for each experiment. We clearly see that our AUC-maximizing variants obtain better classification results than the error-minimizing baseline in most scenarios (and comparable results in few cases). Note that our listwise variant BAM_L obtains better AUROC values than our pairwise variant BAM_P in all scenarios, possibly owing to its tighter lower bound approximation.

4.2 Classification Results on HIV Data Sets

Predicting the effect of a drug using pretreatment genomic information is a current computational challenge in modern medicine. For example, HIV Drug Resistance Database (HIVDB) contains phenotype (i.e., drug susceptibility results) and genotype (i.e., amino acid sequences) information about HIV-1 [14], which is publicly available at http://hivdb.stanford.edu. On HIVDB, we address the problem of predicting drug susceptibility of reverse transcriptase sequences obtained from HIV patients using the genotype information. We extract all sequences originated from subtype B strains and treated with Zalcitabine (known as DDC) or Emtricitabine (known

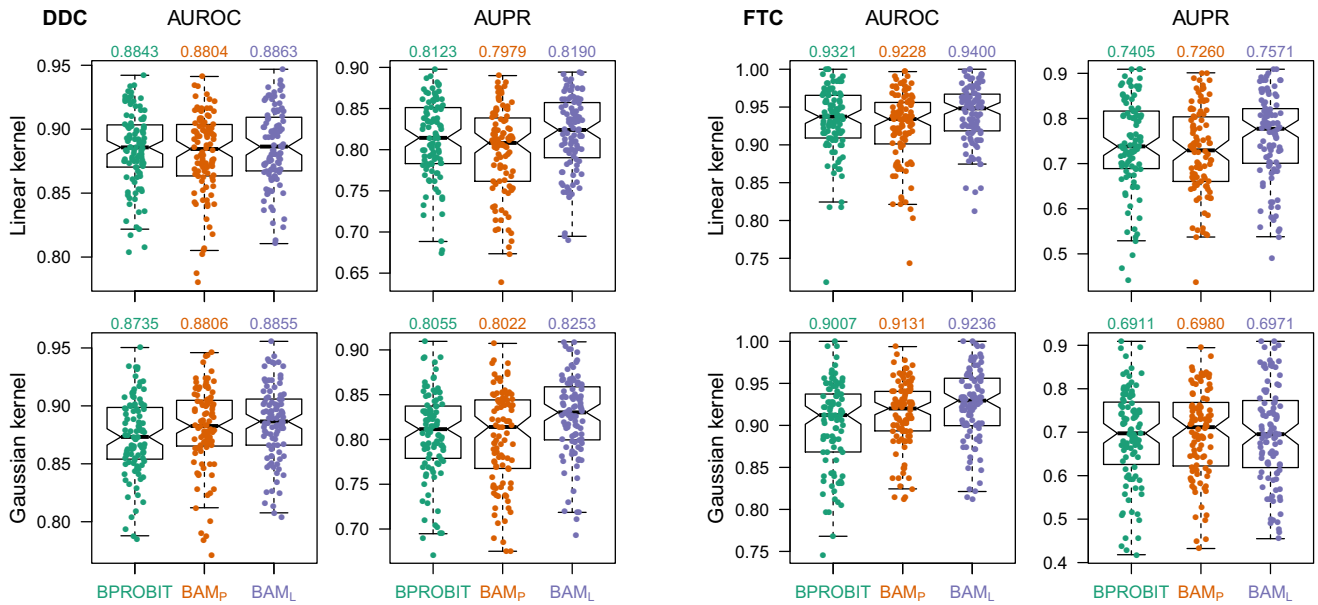

Figure 4. Classification results on two HIV data sets. The box-and-whisker plots show the results of the error-minimizing baseline algorithm (BPROBIT), our AUC-maximizing variant with pairwise ranking (BAM_P), and our AUC-maximizing variant with listwise ranking (BAM_L) over 100 replications in repeated random subsampling validation experiments. The numbers above the figures give the average performance values for each experiment.

as FTC). We remove the sequences with no phenotype or genotype information, leading to two final data sets with 472 (174 susceptible and 298 resistant) and 165 (46 susceptible and 119 resistant) sequences for DDC and FTC, respectively.

We use drug susceptibility results measured using the PhenoSense method for these two nucleoside analogs. Drug susceptibility results are given as fold change:

$$\text{IC}_{50}\text{ ratio} = \frac{\text{IC}_{50}\text{ of an isolate}}{\text{IC}_{50}\text{ of a standard wild-type control isolate}},$$

where IC_{50} of a resistant or wild-type control isolate gives its half maximal inhibitory concentration. We label sequences as "resistant" or "susceptible" using drug-specific cutoff values as done similarly in the earlier studies [8, 15]. The cutoff is set to 1.5 for DDC and to 3.0 for FTC. For each reverse transcriptase, genotype information is extracted from the amino acid sequence of positions 1–240. Amino acid differences from the subtype B consensus wild-type sequence are considered as mutations. There are 856 and 520 unique mutations for DDC and FTC, which means that sequences can be represented as 856- or 520-dimensional binary vectors.

Figure 4 compares the performance of the baseline algorithm and our two variants on two HIV data sets. We see that our listwise variant BAM_L improves AUROC and AUPR values compared to the baseline algorithm in all scenarios, whereas our pairwise variant BAM_P can consistently improve AUROC values only with the Gaussian kernel but not with the linear kernel. Similar to the results on cancer data sets, our listwise variant BAM_L shows better performance than our pairwise variant BAM_P in all scenarios.

5 DISCUSSION

We introduce a novel Bayesian inference framework to optimize AUC measures in Bayesian hierarchical models. This full-Bayesian treatment is made possible by borrowing the pairwise and listwise ranking ideas from the information retrieval literature. To showcase our framework, we develop two linear classification algorithms by modifying Bayesian probit regression model and derive their variational inference procedures. We then illustrate the practical importance of our framework on four biomedical data sets by validation experiments. These results show that our algorithms can obtain better AUROC and AUPR values compared to the baseline error-minimizing algorithm.

To bound the softmax function, we use a simple local variational approximation with a linear Taylor expansion of the log function [2]. An interesting topic for future research is to replace this bound with a much tighter bound such as the one proposed by [12] to further improve the generalization performance of our framework.

REFERENCES

[1] James H. Albert and Siddhartha Chib, 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association*, **88**(422), 669–679, (1993).
[2] David M. Blei and John D. Lafferty, 'A correlated topic model of science', *The Annals of Applied Statistics*, **1**(1), 17–35, (2007).
[3] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, et al., 'Learning to rank using gradient descent', in *Proceedings of the 22nd International Conference on Machine Learning*, (2005).
[4] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li, 'Learning to rank: From pairwise approach to listwise approach', in *Proceedings of the 24th International Conference on Machine Learning*, (2007).
[5] Corinna Cortes and Mehryar Mohri, 'AUC optimization and error rate minimization', in *Advances in Neural Information Processing Systems 16*, (2003).
[6] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer, 'An efficient boosting algorithm for combining preferences', *Journal of Machine Learning Research*, **4**(Nov), 933–969, (2003).
[7] James A. Hanley and Barbara J. McNeil, 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, **143**(1), 29–36, (1982).
[8] Dominik Heider, Robin Senge, Weiwei Cheng, and Eyke Hüllermeier, 'Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction', *Bioinformatics*, **29**(16), 1946–1952, (2013).
[9] Alan Herschtal and Bhavani Raskutti, 'Optimising area under the ROC curve using gradient descent', in *Proceedings of the 21st International Conference on Machine Learning*, (2004).

[10] Thorsten Joachims, 'Optimizing search engines using clickthrough data', in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2002).

[11] Thorsten Joachims, 'A support vector for multivariate performance measures', in *Proceedings of the 22nd International Conference on Machine Learning*, (2005).

[12] David A. Knowles and Tom Minka, 'Non-conjugate variational message passing for multinomial and binary regression', in *Advances in Neural Information Processing Systems 24*, (2011).

[13] Yanyan Lan, Tie-Yan Liu, Zhiming Ma, and Hang Li, 'Generalization analysis of listwise learning-to-rank algorithms', in *Proceedings of the 26th International Conference on Machine Learning*, (2009).

[14] Soo-Yon Rhee, Matthew J. Gonzales, Rami Kantor, Bradley J. Betts, Jaideep Ravela, et al., 'Human immunodeficiency virus reverse transcriptase and protease sequence database', *Nucleic Acids Research*, **31**(1), 298–303, (2003).

[15] Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, Douglas L. Brutlag, et al., 'Genotypic predictors of human immunodeficiency virus type 1 drug resistance', *Proceedings of the National Academy of Sciences of the United States of America*, **103**(46), 17355–17360, (2006).

[16] The Cancer Genome Atlas Network, 'Comprehensive molecular characterization of human colon and rectal cancer', *Nature*, **487**(7407), 330–337, (2012).

[17] The Cancer Genome Atlas Research Network, 'Integrated genomic characterization of endometrial carcinoma', *Nature*, **497**(7447), 67–73, (2013).

[18] Michael E. Tipping, 'Sparse Bayesian learning and the relevance vector machine', *Journal of Machine Learning Research*, **1**(Jun), 211–244, (2001).

[19] Eduardo Vilar and Stephen B. Gruber, 'Microsatellite instability in colorectal cancer — The stable evidence', *Nature Reviews Clinical Oncology*, **7**(3), 153–162, (2010).

[20] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li, 'Listwise approach to learning to rank - Theory and algorithm', in *Proceedings of the 25th International Conference on Machine Learning*, (2008).

[21] Lian Yan, Robert Dodier, Michael C. Mozer, and Richard Wolniewicz, 'Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistics', in *Proceedings of the 20th International Conference on Machine Learning*, (2003).