Stochastic Encodings for Active Feature Acquisition

Alexander Norcliffe, Changhee Lee, Fergus Imrie,
Mihaela van der Schaar, Pietro Liò

Paper ID: 11841

The Problem - Features are not Always Available

- Active Feature Acquisition (AFA): Sequentially select what to measure to improve *long term* predictive power, based on *existing, instance-wise* information

The Problem - Features are not Always Available

- Active Feature Acquisition (AFA): Sequentially select what to measure to improve *long term* predictive power, based on *existing, instance-wise* information
- Application: Doctor diagnosing a patient, they choose the test based on current observations for each individual patient

Existing Approaches

- Reinforcement Learning (RL)
  - Natural solution for sequential decision making
  - Suffers from training difficulty

$$\underset{i \in [d] \setminus O}{\mathbf{argmax}}\ \pi_\theta(\mathbf{x}_O)_i$$

Existing Approaches

- Reinforcement Learning (RL)
  - Natural solution for sequential decision making
  - Suffers from training difficulty
- Maximize Conditional Mutual Information
  - Grounded in information theory
  - Makes myopic acquisitions
  - Can be maximized by eliminating options

$$\underset{i \in [d] \setminus O}{\mathbf{argmax}} \; \pi_\theta(\mathbf{x}_O)_i$$

$$\underset{i \in [d] \setminus O}{\mathbf{argmax}} \; I(X_i; Y | \mathbf{x}_O)$$

Indicator Example - CMI is Myopic

Binary classification, one feature is "The Indicator", telling us which feature gives the label:

$$\mathbf{x} = [0, 0, 1, 0, 1, 3], \qquad y = 1$$

Indicator Example - CMI is Myopic

Binary classification, one feature is "The Indicator", telling us which feature gives the label:

$$\mathbf{x} = [0, 0, \boxed{1}, 0, 1, \boxed{3}], \qquad y = 1$$

Indicator Example - CMI is Myopic

Binary classification, one feature is "The Indicator", telling us which feature gives the label:

$$\mathbf{x} = [0, 0, \boxed{1}, 0, 1, \boxed{3}], \qquad y = 1$$

CMI optimizes for immediate predictive power - does not select indicator first

Indicator Example - CMI is Myopic

Binary classification, one feature is "The Indicator", telling us which feature gives the label:

$$\mathbf{x} = [0, 0, \boxed{1}, 0, 1, \boxed{3}], \qquad y = 1$$

CMI optimizes for immediate predictive power - does not select indicator first

Insight: Considering possible values of unobserved features is *necessary* for optimality and can be *sufficient*:

$$\underset{i \in [d] \setminus O}{\operatorname{argmax}} \; \mathbb{E}_{p(\mathbf{x}_U | \mathbf{x}_O)} I(X_i ; Y | \mathbf{x}_U, \mathbf{x}_O)$$

Entropy Example

- CMI maximization can be achieved by making low likelihoods lower:

$$H([0.5, 0.5, 0.0]) = 0.693$$

$$H([0.7, 0.15, 0.15]) = 0.819$$

Entropy Example

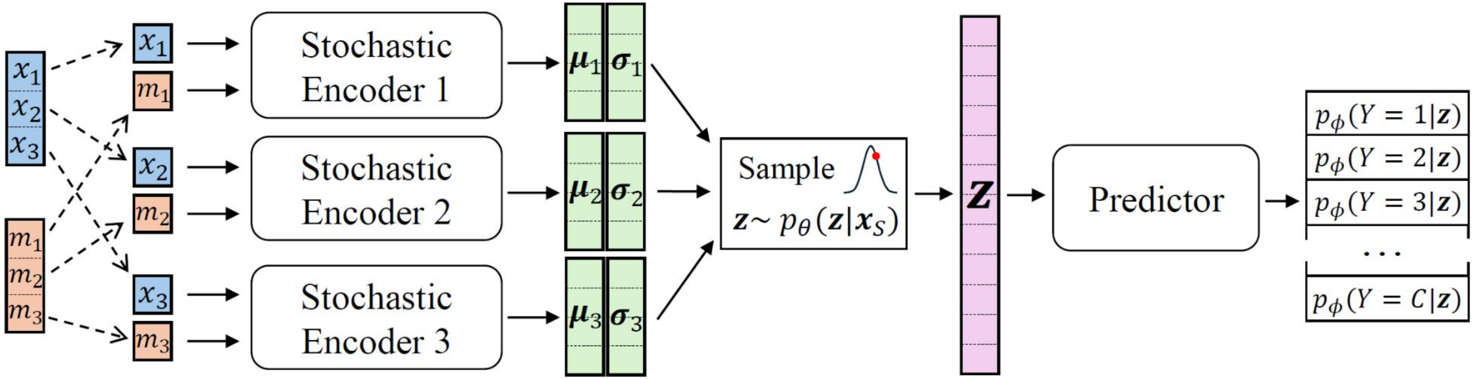
- CMI maximization can be achieved by making low likelihoods lower:

$$H([0.5, 0.5, 0.0]) = 0.693$$

$$H([0.7, 0.15, 0.15]) = 0.819$$

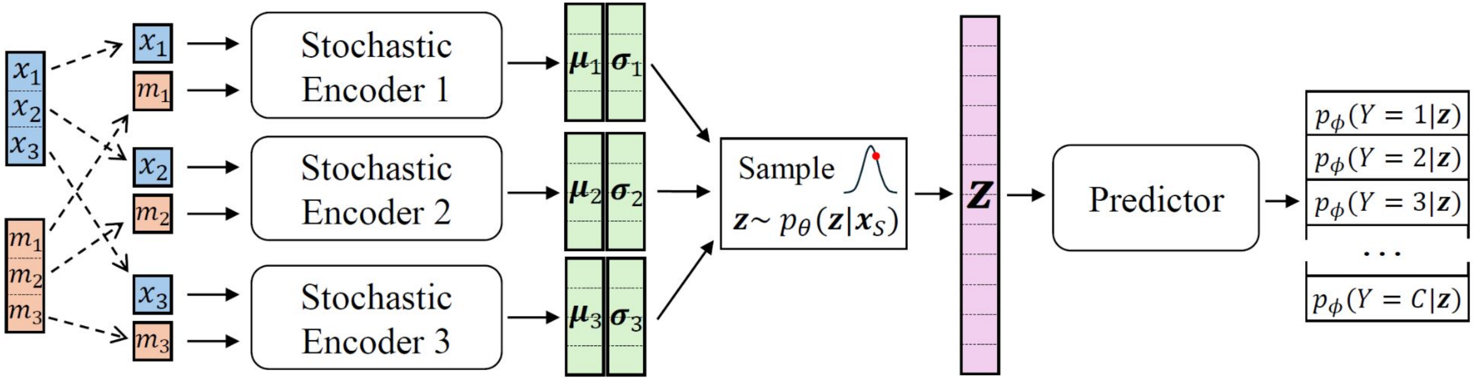
- Focus should be placed on identifying the most likely class, not on confirming which ones are incorrect

SEFA- Architecture



- Each feature is separately encoded

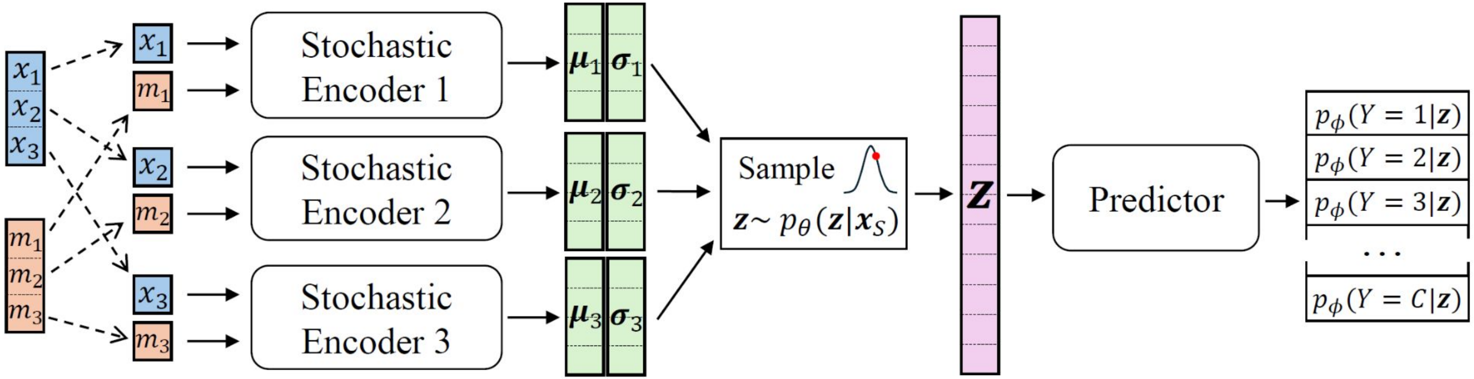
SEFA- Architecture



- Each feature is separately encoded
- Predictions made on latent samples, multiple samples are taken to make full prediction

$$p_{\theta,\phi}(y|\mathbf{x}_S) = \mathop{\mathbb{E}}_{p_\theta(\mathbf{z}|\mathbf{x}_S)} p_\phi(y|\mathbf{z})$$

SEFA- Architecture



- Each feature is separately encoded
- Predictions made on latent samples, multiple samples are taken to make full prediction
- Supervised training with negative log-likelihood and information bottleneck regularization - avoids RL training

$$p_{\theta,\phi}(y|\mathbf{x}_S) = \mathop{\mathbb{E}}_{p_\theta(\mathbf{z}|\mathbf{x}_S)} p_\phi(y|\mathbf{z})$$

$$L = -\log p_{\theta,\phi}(Y|X_S) + \beta I_\theta(Z; X_S)$$

SEFA - Acquisition Objective

$$\underset{i \in [d] \setminus O}{\operatorname{argmax}} \sum_{c \in [C]} p_{\theta,\phi}(Y = c | \mathbf{x}_O) \underset{p_\theta(\mathbf{z}|\mathbf{x}_O)}{\mathbb{E}} r(c, \mathbf{z}, i)$$

SEFA - Acquisition Objective

$$\underset{i \in [d] \setminus O}{\operatorname{argmax}} \sum_{c \in [C]} p_{\theta,\phi}(Y = c | \mathbf{x}_O) \underset{p_\theta(\mathbf{z}|\mathbf{x}_O)}{\mathbb{E}} r(c, \mathbf{z}, i)$$

<u>Latent Gradients as Importance Measure:</u>

$$r(c, \mathbf{z}, i) = \frac{||\mathbf{g}_{\mathcal{G}_i}||_2}{\sum_j ||\mathbf{g}_{\mathcal{G}_j}||_2}$$

$$\mathbf{g} = \nabla_{\mathbf{z}} p_\phi(Y = c|\mathbf{z})$$

Gradients measure importance of latents, aggregated across the feature that encodes them

SEFA - Acquisition Objective

$$\underset{i\in[d]\setminus O}{\mathrm{argmax}} \sum_{c\in[C]} p_{\theta,\phi}(Y=c|\mathbf{x}_O) \underset{p_\theta(\mathbf{z}|\mathbf{x}_O)}{\mathbb{E}} r(c,\mathbf{z},i)$$

<u>Stochastic Encoders:</u>
Consider many possible unobserved feature values in current decision

<u>Latent Gradients as Importance Measure:</u>

$$r(c,\mathbf{z},i) = \frac{||\mathbf{g}_{\mathcal{G}_i}||_2}{\sum_j ||\mathbf{g}_{\mathcal{G}_j}||_2}$$

$$\mathbf{g} = \nabla_{\mathbf{z}} p_\phi(Y=c|\mathbf{z})$$

Gradients measure importance of latents, aggregated across the feature that encodes them

SEFA - Acquisition Objective

$$\underset{i \in [d] \setminus O}{\operatorname{argmax}} \sum_{c \in [C]} p_{\theta,\phi}(Y = c | \mathbf{x}_O) \underset{p_\theta(\mathbf{z}|\mathbf{x}_O)}{\mathbb{E}} r(c, \mathbf{z}, i)$$

<u>Probability Weighting:</u>
Place more focus on distinguishing between likely labels

<u>Stochastic Encoders:</u>
Consider many possible unobserved feature values in current decision

<u>Latent Gradients as Importance Measure:</u>

$$r(c, \mathbf{z}, i) = \frac{||\mathbf{g}_{\mathcal{G}_i}||_2}{\sum_j ||\mathbf{g}_{\mathcal{G}_j}||_2}$$

$$\mathbf{g} = \nabla_{\mathbf{z}} p_\phi(Y = c | \mathbf{z})$$

Gradients measure importance of latents, aggregated across the feature that encodes them

Why use the Latent Space?

$$\underset{i \in [d] \setminus O}{\operatorname{argmax}} \sum_{c \in [C]} p_{\theta,\phi}(Y = c | \mathbf{x}_O) \underset{p_\theta(\boxed{\mathbf{z}} | \mathbf{x}_O)}{\mathbb{E}} r(c, \boxed{\mathbf{z}}, i)$$

Why use the Latent Space?

$$\underset{i \in [d] \setminus O}{\operatorname{argmax}} \sum_{c \in [C]} p_{\theta,\phi}(Y = c | \mathbf{x}_O) \mathop{\mathbb{E}}_{p_\theta(\boxed{\mathbf{z}} | \mathbf{x}_O)} r(c, \boxed{\mathbf{z}}, i)$$

- Gradients are more meaningful and comparable for the latents (same scale, all continuous)

Why use the Latent Space?

$$\underset{i \in [d] \setminus O}{\text{argmax}} \sum_{c \in [C]} p_{\theta,\phi}(Y = c | \mathbf{x}_O) \underset{p_\theta(\boxed{\mathbf{z}} | \mathbf{x}_O)}{\mathbb{E}} r(c, \boxed{\mathbf{z}}, i)$$

- Gradients are more meaningful and comparable for the latents (same scale, all continuous)
- Latents have less noise

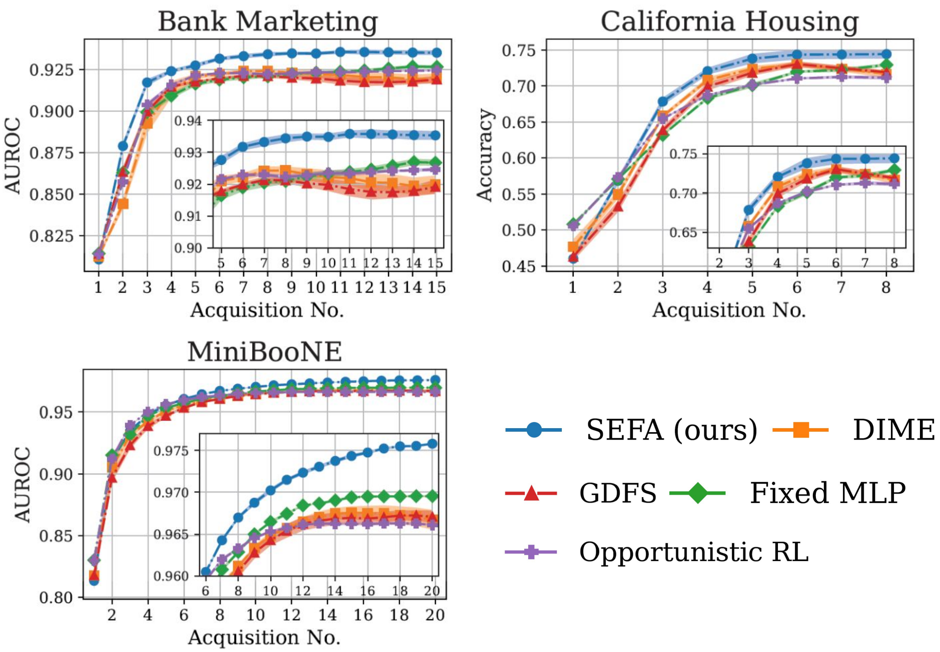
Why use the Latent Space?

$$\underset{i \in [d] \setminus O}{\operatorname{argmax}} \sum_{c \in [C]} p_{\theta,\phi}(Y = c | \mathbf{x}_O) \mathop{\mathbb{E}}_{p_\theta(\mathbf{z} | \mathbf{x}_O)} r(c, \mathbf{z}, i)$$

- Gradients are more meaningful and comparable for the latents (same scale, all continuous)
- Latents have less noise
- Do not need to learn complex generative model

Results - Tabular Data

Model	Bank Marketing	California Housing	MiniBooNE
DIME	0.907 ± 0.002	0.661 ± 0.002	0.951 ± 0.001
Fixed MLP	0.909 ± 0.001	0.658 ± 0.002	0.954 ± 0.000
GDFS	0.907 ± 0.001	0.653 ± 0.002	0.949 ± 0.000
ORL	0.910 ± 0.000	0.657 ± 0.001	0.953 ± 0.000
SEFA	**0.919 ± 0.001**	**0.676 ± 0.005**	**0.957 ± 0.000**

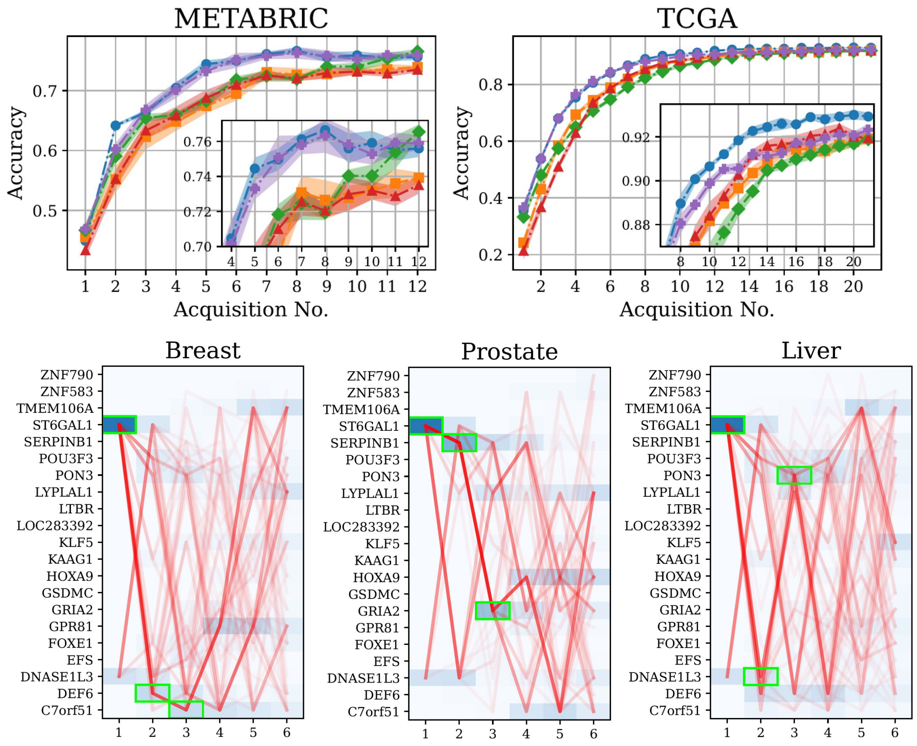


Stochastic Encodings for
Active Feature Acquisition

Paper ID: 11841

Results - Cancer Classification

Model	METABRIC	TCGA
DIME	0.670 ± 0.007	0.805 ± 0.002
Fixed MLP	0.685 ± 0.003	0.799 ± 0.004
GDFS	0.671 ± 0.005	0.797 ± 0.002
ORL	0.706 ± 0.004	0.838 ± 0.002
SEFA	**0.709 ± 0.003**	**0.843 ± 0.002**



Stochastic Encodings for
Active Feature Acquisition

Paper ID: 11841