Functional Bregman Divergence, Bayesian Estimation of Distributions and Completely Lazy Classifiers

Maya R. Gupta
Dept. of Electrical Engineering
University of Washington

Santosh Srivastava
Fred Hutch Cancer Research Center

Bela A. Frigyik
Dept. of Mathematics
Purdue University

Sergey Feldman
Dept. of Electrical Engineering
University of Washington

Eric Garcia
Dept. of Electrical Engineering
University of Washington
Bregman divergence

Functional Bregman divergence

Bayesian estimation of distributions

- Uniform
- Gaussian

Bayesian QDA

- Local BDA
- Gaussian mixture

completely lazy learning
The mean minimizes average squared error

Let \( x_1, x_2, \ldots, x_N \in \mathbb{R}^n \).

\[
A^* = \arg \min_{A \in \mathbb{R}^n} \frac{1}{N} \sum_j (\|x_j - A\|_2^2)
\]

Then,

\[
A^* = \frac{1}{N} \sum_j x_j
\]
The mean minimizes average squared error

Let \( x_1, x_2, \ldots, x_N \in \mathbb{R}^n \).

\[
A^* = \arg \min_{A \in \mathbb{R}^n} \frac{1}{N} \sum_j (\|x_j - A\|_2^2)
\]

Then,

\[
A^* = \frac{1}{N} \sum_j x_j
\]

Are there other distortion functions that yield the sample mean?
The mean minimizes average Bregman divergence  
(Banerjee et al. JMLR ’05, IEEE Trans. on Info Theory ’05)

Let \( x_1, x_2, \ldots, x_N \in \mathbb{R}^n \).

Let \( d(x, y) \) be any Bregman divergence.

\[
A^* = \arg \min_{A \in \mathbb{R}^n} \frac{1}{N} \sum_j d(x_j, A)
\]

Then,

\[
A^* = \frac{1}{N} \sum_j x_j
\]
Bregman divergence between vectors

Class of distortion functions, including:

- sum of squared errors
- relative entropy
- Itakura-Saito distance
- etc.

General formula:

\[
d_\phi(x, y) = \phi(x) - \phi(y) - \nabla \phi(y)^T(x - y), \quad x, y \in \mathbb{R}^n
\]

\(\phi\) is convex function.

Total squared error:

\[
\phi(x) = \sum_i x[i]^2
\]

Relative entropy:

\[
\phi(x) = \sum_i x[i] \log x[i]
\]
Class of distortion functions, including:

- sum of squared errors
- relative entropy
- Itakura-Saito distance
- etc.

General formula:
\[ d_\phi(x, y) = \phi(x) - \phi(y) - \nabla \phi(y)^T (x - y), \quad x, y \in \mathbb{R}^n \]

\( \phi \) is convex function.

\( d_\phi(x, y) \) is tail of Taylor series expansion of \( \phi \) around \( y \):
\[ \phi(x) = \phi(y) + \nabla \phi(y)^T (x - y) + d_\phi(x, y) \]
Relationship to Bayesian Estimation

Let $d(x, y)$ be any Bregman divergence.

**Goal:** Estimate a parameter $\hat{\theta} \in \mathbb{R}$,
Given candidates $\theta \in \mathbb{R}$ and posterior $p(\theta)$.

Consider the **Bayesian estimate** with $d$ as the risk function:

$$\theta^* = \arg\min_{\hat{\theta} \in \mathbb{R}} \int_{\theta} p(\theta) d(\theta, \hat{\theta}) d\theta = \arg\min_{\hat{\theta} \in \mathbb{R}} E_{\Theta}[d(\Theta, \hat{\theta})]$$
Relationship to Bayesian Estimation

Let \( d(x, y) \) be any Bregman divergence.

**Goal:** Estimate a parameter \( \hat{\theta} \in \mathbb{R} \),
Given candidates \( \theta \in \mathbb{R} \) and posterior \( p(\theta) \).

Consider the **Bayesian estimate** with \( d \) as the risk function:

\[
\theta^* = \arg \min_{\hat{\theta} \in \mathbb{R}} \int_{\theta} p(\theta) d(\theta, \hat{\theta}) d\theta = \arg \min_{\hat{\theta} \in \mathbb{R}} E_{\Theta}[d(\Theta, \hat{\theta})]
\]

Say you flip
1 tail
9 heads
Let \( \theta = P(\text{tails}) \)

![Graph of likelihood of \( \theta \) given 1 tail and 9 heads](image)
Relationship to Bayesian Estimation

Let $d(x, y)$ be any Bregman divergence.

**Goal:** Estimate a parameter $\hat{\theta} \in \mathbb{R}$,

Given candidates $\theta \in \mathbb{R}$ and posterior $p(\theta)$.

Consider the **Bayesian estimate** with $d$ as the risk function:

$$\theta^* = \arg \min_{\hat{\theta} \in \mathbb{R}} \int_{\theta} p(\theta) d(\theta, \hat{\theta}) d\theta = \arg \min_{\hat{\theta} \in \mathbb{R}} E_\Theta[d(\Theta, \hat{\theta})]$$

Then, Banerjee et al. theorem says minimizer is the mean:

$$\theta^* = E_\Theta[\Theta]$$

<table>
<thead>
<tr>
<th>Say you flip 1 tail</th>
<th>(\theta^* = .1666)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 heads</td>
<td>(\hat{\theta}_{MLE} = .1)</td>
</tr>
<tr>
<td>Let (\theta = P(tails))</td>
<td></td>
</tr>
</tbody>
</table>
Estimation of Distributions

**Goal:** Estimate a distribution $\hat{f}(x)$
Given candidates $f : \mathbb{R} \to \mathbb{R}$, and posterior $p(f)$.

**Ex:** Given samples $\{2, 3, 7, 8\}$, estimate the generating uniform distribution $U[0, a]$. 
Estimation of Distributions

**Goal:** Estimate a distribution $\hat{f}(x)$
Given candidates $f : \mathbb{R} \rightarrow \mathbb{R}$, and posterior $p(f)$.

**Ex:** Given samples $\{2, 3, 7, 8\}$, estimate the generating uniform distribution $U[0, a]$.

$$\hat{f}(x)_{ML} = U[0, 8]$$
**Estimation of Distributions**

**Goal:** Estimate a distribution $\hat{f}(x)$

Given candidates $f : \mathbb{R} \rightarrow \mathbb{R}$, and posterior $p(f)$.

**Ex:** Given samples $\{2, 3, 7, 8\}$, estimate the generating uniform distribution $U[0, a]$.

Bayesian parameter estimate:

$$\arg\min_{\hat{a} \in \mathbb{R}^+} E_A[(A - \hat{a})^2]$$
Estimation of Distributions

**Goal:** Estimate a distribution \( \hat{f}(x) \)
Given candidates \( f : \mathbb{R} \to \mathbb{R} \), and posterior \( p(f) \).

**Ex:** Given samples \{2, 3, 7, 8\}, estimate the generating uniform distribution \( U[0, a] \).

Bayesian parameter estimate:

\[
\arg \min_{\hat{a} \in \mathbb{R}^+} E_A[(A - \hat{a})^2]
\]

Assume gamma prior for \( A \):

\[
\Rightarrow U \left[ 0, \frac{1}{\gamma_0} \frac{P(\chi^2_{\gamma_1} < \frac{2}{\gamma_2 X_{\text{max}}})}{P(\chi^2_{\gamma_3} < \frac{2}{\gamma_4 X_{\text{max}}})} \right]
\]
Bayesian Estimation of Distributions (Frigyik, Gupta, Srivastava ’06)

**Goal:** Estimate a distribution $\hat{f}(x)$

Given candidates $f : \mathbb{R} \to \mathbb{R}$, and posterior $p(f)$.

**Ex:** Given samples $\{2, 3, 7, 8\}$, estimate the generating uniform distribution $U[0, a]$.

$$f^*(x) = \arg\min_{\hat{f}(x)} EF[d(F, \hat{f})]$$

$$\equiv \arg\min_{\hat{f}(x)} \int_{f \in U} d(f, \hat{f})p(f)dU$$
Bayesian Estimation of Distributions (Gupta and Srivastava ’06)

Goal: Estimate a distribution $\hat{f}(x)$
Given candidates $f : \mathbb{R} \rightarrow \mathbb{R}$, and posterior $p(f)$.

**Ex:** Given samples $\{2, 3, 7, 8\}$, estimate the generating uniform distribution $U[0, a]$.

$$f^*(x) = \arg \min_{\hat{f}(x)} EF[d(F, \hat{f})]$$

$$\equiv EF[F]$$

if $d$ is a Bregman divergence?

What is a Bregman divergence between functions?
Bregman Divergence Definitions

**Bregman Divergence:** for vectors $x, y \in \mathbb{R}^n$, convex function $\phi$, 
$$d_\phi(x, y) = \phi(x) - \phi(y) - \nabla\phi(y)^T(x - y),$$

**Pointwise Bregman Divergence:** for functions $f(t), g(t)$ 
$$d_\phi(f, g) = \int_t d_\phi(f(t), g(t))d\nu(t),$$

$$(\text{Jones and Byrne 1990, Csiszar 1995})$$

$$\arg\min_{\hat{f}(x)} E_F[d(F, \hat{f})] = E_F[F]$$
Bregman Divergence Definitions

**Bregman Divergence:** for vectors $x, y \in \mathbb{R}^n$, convex function $\phi$, 
$$d_\phi(x, y) = \phi(x) - \phi(y) - \nabla \phi(y)^T (x - y),$$

**Pointwise Bregman Divergence:** for functions $f(t), g(t)$ 
$$d_\phi(f, g) = \int_t d_\phi(f(t), g(t)) d\nu(t),$$

*(Jones and Byrne 1990, Csiszar 1995)*

**Functional Bregman Divergence:** *(Srivastava, Gupta, Frigyik, JMLR 06)* 
$f, g : \mathbb{R}^n \to \mathbb{R}$ and $f, g \geq 0$, and $f, g \in L^p(\nu)$ 
$\phi : L^p(\nu) \to \mathbb{R}$, strictly convex functional, $\phi \in C^2$

$$d_\phi(f, g) = \phi[f] - \phi[g] - \delta\phi[g; f - g]$$ 
Frechet derivative of $\phi$ at $g$ in the direction of $f - g$
Functional Bregman Divergence


\[ f, g : \mathbb{R}^n \to \mathbb{R} \text{ and } f, g \geq 0, \text{ and } f, g \in L^p(\nu) \]
\[ \phi : L^p(\nu) \to \mathbb{R}, \text{ strictly convex functional, } \phi \in C^2 \]

\[ d_\phi(f, g) = \phi[f] - \phi[g] - \delta \phi[g; f - g] \]

Frechet derivative of \( \phi \)

at \( g \) in the direction of \( f - g \)

Frechet derivative:

\[ \phi[g + a] - \phi[g] = \delta \phi[g; a] + \epsilon[g, a]\|a\|_{L^p(\nu)} \]

For all \( a \in L^p(\nu) \), with \( \epsilon[g, a] \to 0 \) as \( \|a\|_{L^p(\nu)} \to 0 \).
Functional Bregman Divergence

\( \phi : L^p(\nu) \to \mathbb{R} \), strictly convex functional, \( \phi \in C^2 \)

\[
d_\phi(f, g) = \phi[f] - \phi[g] - \delta \phi[g; f - g]
\]

Frechet derivative of \( \phi \)
at \( g \) in the direction of \( f - g \)

**Ex:** total squared error \( \phi[g] = \int g^2 d\nu \)

Compare with vector Bregman divergence
\[
\phi(x) = \sum_i x[i]^2
\]
Functional Bregman Divergence


\[ f, g : \mathbb{R}^n \to \mathbb{R}, \ f, g, \in L^p(\nu) \]
\[ \phi : L^p(\nu) \to \mathbb{R}, \text{ strictly convex functional, } \phi \in C^2 \]

\[
d_\phi(f, g) = \phi[f] - \phi[g] - \delta\phi[g; f - g]
\]
Frechet derivative of \( \phi \)
at \( g \) in the direction of \( f - g \)

**Ex:** total squared error \( \phi[g] = \int g^2 d\nu \)

\[ \Rightarrow \delta\phi[g; f - g] = \int 2g(f - g) d\nu \]

\[
d_\phi(f, g) = \int f^2 d\nu - \int g^2 d\nu - \int 2g(f - g) d\nu
\]

\[ = \int (f - g)^2 d\nu = \| f - g \|_{L^2(\nu)}^2 \]
Functional Bregman Divergence


\[ f, g : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f, g, \in L^p(\nu) \]
\[ \phi : L^p(\nu) \rightarrow \mathbb{R}, \text{ strictly convex functional, } \phi \in C^2 \]

\[
d_\phi(f, g) = \phi[f] - \phi[g] - \delta\phi[g; f - g]
\]

Frechet derivative of \( \phi \) at \( g \) in the direction of \( f - g \)

Functional Bregman divergences include pointwise Bregman divergences and more!

Ex: squared bias

\[
d_\phi(f, g) = \left( \int (f - g) d\nu \right)^2 \quad \phi[g] = \left( \int g d\nu \right)^2
\]
Functional Bregman divergence has same properties as Bregman divergence

Non-negativity
Convexity with respect to first function
Linearity with respect to $\phi$-functionals
Equivalence classes with respect to $\phi$-functionals
Dual divergences by Legendre transformation
Generalized Pythagorean Inequality
Theorem: for random function $F$ defined on a finite-dimensional manifold with posterior $p_F$,

$$f^* = \arg \min_{\hat{f}} E_F[d_\phi[F, \hat{f}]]$$

$$\equiv E_F[F]$$
Bayesian Estimation of Distributions


Theorem: for random function $F$ defined on a finite-dimensional manifold with posterior $p_F$,

$$f^* = \arg \min_{\hat{f}} E_F[d_\phi[F, \hat{f}]]$$

$$\equiv E_F[F]$$

e.g. parametric distribution or decomposable in terms of finite basis functions
Uniform Example (arXiv: Frigyik, Srivastava, Gupta)

Ex: Given samples \{2, 3, 7, 8\}, estimate the generating uniform distribution \(U[0, a]\).

Let \(F\) be a random uniform distribution: \(U[0, a]\)
Let \(p_F\) be the likelihood of \(F\) given \(N\) data samples.

\[
f^* = \arg \min_{\hat{f}} E_F[d_\phi[F, \hat{f}]]
\]

\[
\equiv E_F[F]
\]

\[
f^*(x) = \frac{\int_a=\max(x,X_{\text{max}}) \frac{1}{a} \left(\frac{1}{a^N}\right) \left\| \frac{df}{da} \right\|_2 ^2 da}{\int_a=X_{\text{max}} \frac{1}{a^N} \left\| \frac{df}{da} \right\|_2 ^2 da}
\]

(actually, we use the Fisher information metric for \(d\mathcal{U}\))
Bayesian Estimation of Distributions \textit{(arXiv: Frigyik, Srivastava, Gupta)}

Ex: Given samples \{2, 3, 7, 8\}, estimate the generating uniform distribution $U[0, a]$.

Let $F$ be a random uniform distribution: $U[0, a]$
Let $p_F$ be the likelihood of $F$ given $N$ data samples.

\[
f^* = \arg\min_{\hat{f}} E_F[d_\phi[F, \hat{f}]]
\equiv E_F[F]
\]

\[
f^*(x) = \frac{\int_{\max(x, X_{max})}^{\infty} \frac{1}{a} \left( \frac{1}{a^N} \right) \frac{da}{a}}{\int_{X_{max}}^{\infty} \frac{1}{a^N} \frac{da}{a}}
\]
Bayesian Estimation of Distributions (arXiv: Frigyik, Srivastava, Gupta)

\[
f^*(x)_{\text{Bayesian}} = \arg \min_{\hat{f}(x)} E_F[d_\phi[F, \hat{f}]]
\]

\[
\equiv E_F[F]
\]

\[
= \frac{N(X_{\text{max}})^N}{(N + 1)[\max(x, X_{\text{max}})]^{N+1}}
\]
Compare estimates

Let $F$ be a random uniform distribution: $U[0,a]$
Let $p_F$ be the likelihood of $F$ given the data samples.

$$\arg \min_{\hat{f}(x)} E_F[d_\phi[F, \hat{f}]] = \frac{N(X_{max})^N}{(N + 1)[\max(x, X_{max})]^{N+1}}$$

$$\arg \min_{\hat{f}(x) \in U} E_F[(\|F - \hat{f}\|_2)^2] = U[0, X_{max}2^{1/N}]$$

Bayesian parameter estimate of $a$, gamma prior $p(a)$:

$$\arg \min_{\hat{a} \in \mathbb{R}^+} E_A[(A - \hat{a})^2] \Rightarrow U \left[0, \frac{1}{\gamma_0} \frac{P(\chi_{\gamma_1}^2 < \frac{2}{\gamma_2 X_{max}^2})}{P(\chi_{\gamma_3}^2 < \frac{2}{\gamma_4 X_{max}^2})} \right]$$
Compare estimates

Simulation: Draw $n$ random samples from $U[0, 1]$
Metric: Squared error between estimated dist. and $U[0, 1]$. 

![Graph showing comparison of estimates with respect to number of data samples.](image-url)
Bregman divergence

Functional Bregman divergence

Bayesian estimation of distributions

Uniform

Gaussian

Bayesian QDA

Local BDA

Gaussian mixture

completely lazy learning
Classification set-up

Training Data $T = \{ X_i, Y_i \}$
Feature vectors $X_i \in \mathbb{R}^d$ for $i = 1, \ldots n$.
Associated labels $Y_i \in \mathcal{G}$, where $\mathcal{G}$ is a finite set of classes.

Test vector $X$, estimates its associated label $\hat{Y}$. 
QDA: classifying with Gaussian models
QDA: classifying with Gaussian models
QDA: classifying with Gaussian models
Bayesian: Minimizing Expected Misclassification Costs

\[ Y = \arg \min_g \sum_{h=1}^G C(g, h) p(x|Y = h) P(Y = h) \]

\[ \hat{Y} = \arg \min_g E \left[ \sum_{h=1}^G C(g, h) N_h(x) \Theta_h \right] \]

\[ \equiv \arg \min_g \sum_{h=1}^G C(g, h) E[N_h(x)] E_{\Theta} [\Theta_h] \]

\[ E_{\mu_h, \Sigma_h} [N_h(x)] \quad E_{N_h} [N_h(x)] \]

(Geisser 1964)

(Srivastava, Gupta 2006)
Distribution-based Bayesian Minimum Expected Misclassification Cost:
(Srivastava and Gupta, IEEE ISIT (2006))

For a test point $x$ and class $h$,

$$E_{N_h}[N_h(x)] = \int_M N(x) f(N|T_h) dM$$

- Look at all possible Gaussians
- Prob. of test point given some Gaussian
- Prob. of that Gaussian given training data and prior
- Measure over space of Gaussians

$$dM = \frac{d\mu d\Sigma}{|\Sigma|^{d+2}}$$

differential element based on Fisher information matrix (C. R. Rao ’45).
Distribution-based Bayesian Minimum Expected Misclassification Cost:

\[(Srivastava and Gupta, IEEE ISIT (2006))\]

For a test point \(x\) and class \(h\),

\[E_{N_h}[N_h(x)] = \int M \mathcal{N}(x) f(N|T_h) dM\]

- Prob. of a Gaussian given training data
- \(f(N|T_h) = \prod_{j=1}^{k} \mathcal{N}(X_j)p(N)\)
  - Likelihood of the iid training samples
  - Prior prob of that Gaussian
Prior matters with minimum expected risk

Design goals for the prior (over the Gaussian distributions):

1) Regularize for ill-posed likelihood to reduce estimation variance (not a flat prior).

2) Add sensible bias.

3) Allow the estimation to converge as number of training samples becomes infinite.

4) Lead to closed form solution.
Proposed Prior

\[
p(N_h) = \gamma_0 \exp\left(-\frac{1}{2} \text{trace}(\Sigma_h^{-1} B_h)\right) \left| \Sigma_h \right|^{-\frac{q}{2}} \quad \text{(inverted Wishart)}
\]

\[
\Sigma_{h,max} = \frac{B_h}{q}
\]

We set:

\[
B_h = q \text{ diag}(\hat{\Sigma}_{h,ML})
\]
Proposed Prior

\[ p(\mathcal{N}_h) = \gamma_0 \frac{\exp\left(-\frac{1}{2} \text{trace}(\Sigma_h^{-1} B_h)\right)}{|\Sigma_h|^\frac{q}{2}} \]  

(involved Wishart)
Distribution-based classifier and closed form solution

Choose the class \( \hat{Y} = g \in G \) that minimizes

\[
\sum_{h=1}^{G} C(g, h) E_{N_h}[N_h(x)] E_{\Theta}[\Theta_h]
\]

Closed-form solution:

\[
E_{N_h}[N_h(x)] = \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)(1 + \frac{n_h}{n_h+1} Z_h^T D_h^{-1} Z_h)^{-\frac{n_h+q+1}{2}}}{\pi^{\frac{d}{2}} \Gamma\left(\frac{n_h+q-d+1}{2}\right)|\left(\frac{n_h+1}{n_h}\right)D_h|^\frac{1}{2}}
\]

\[
B_h = .95q \text{ diag}(\hat{\Sigma}_{ML,h}) + .05I
\]

\[
D_h = S_h + B_h, \text{ and } Z_h = x - \bar{x}_h
\]
Distribution-based Bayesian discriminant (Srivastava, Gupta, 2006)

\[ E_{N_h}[N_h] = \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)(1+\frac{n_h}{n_h+1}Z_h^T D_h^{-1} Z_h)^{-\frac{n_h+q+1}{2}}}{\pi^\frac{d}{2} \Gamma\left(\frac{n_h+q-d+1}{2}\right)\left|\frac{n_h+1}{n_h}D_h\right|^\frac{1}{2}} \]

Parameter-based Bayesian discriminant (Geisser, 1964)

\[ E_{\mu,\Sigma}[N_h] = \frac{\Gamma\left(\frac{n_h+q-d-1}{2}\right)(1+\frac{n_h}{n_h+1}Z_h^T D_h^{-1} Z_h)^{-\frac{n_h+q-d-1}{2}}}{\pi^\frac{d}{2} \left|\frac{n_h+1}{n_h}D_h\right|^\frac{1}{2} \Gamma\left(\frac{n_h+q-2d-1}{2}\right)} \]

Difference: For parameter-based you need \( n_h > 2d - q + 1 \) samples for each class. If you have few samples, forced to use high \( q = \) more bias.
Bregman divergence

Functional Bregman divergence

Bayesian estimation of distributions

Uniform

Gaussian

Bayesian QDA

(reduce model bias)

Local BDA

Gaussian mixture

completely lazy learning
Local BDA

1. Find the $k$ samples from each class nearest to the test sample.
2. Fit a Gaussian to the nearest $k$ samples of each class.
3. Classify as the class that minimizes expected misclassification costs.

Related Work:
Local Nearest Means (Mitani and Hamamoto, 2000)
SVM-KNN (Malik et al. 2006)
Local BDA – 7 Neighbors
Local BDA – 7 Neighbors

Feature 2

Feature 1

$X[2]$
Local BDA- 7 neighbors

Feature 2

\( X[2] \)

Feature 1

\( X[1] \)
How do we choose the neighborhood size?


- Theoretically-sound only if training and test are iid.
- If training set evolving, must re-train
How do we choose the neighborhood size?

Standard: cross-validate on training. Not very lazy.

Proposed: average over multiple neighborhood sizes = completely lazy.

Choose the class \( \hat{Y} \) that solves

\[
\arg \min_{g=1,\ldots,G} \sum_{h=1}^{G} C(g, h) \underbrace{E_{N_{h,K}[N_{h,K}(x)]}E_{\Theta}[\Theta_h]}_{\text{average discriminant with respect to uncertainty in the Gaussian-fit to the training samples and to the neighborhood size}}
\]
## Representative Misclassification Rates on UCI datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>Optical Char. Rec.</th>
<th>Isolet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cv k</td>
<td>E_K[]</td>
</tr>
<tr>
<td>Local Nearest Means</td>
<td>3.3</td>
<td>3.2</td>
</tr>
<tr>
<td>Local BDA</td>
<td>2.6</td>
<td>1.7</td>
</tr>
<tr>
<td>kNN</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>DANN</td>
<td>4.0</td>
<td>4.3</td>
</tr>
<tr>
<td>SVM-kNN</td>
<td>2.6</td>
<td>2.7</td>
</tr>
<tr>
<td>SVM</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>GMM</td>
<td>10.9</td>
<td></td>
</tr>
<tr>
<td>GMM BQDA</td>
<td>5.5</td>
<td></td>
</tr>
</tbody>
</table>
Summary

1. Proposed a functional Bregman divergence for functions $f, g$:

   \[ d_{\phi}(f, g) = \phi[f] - \phi[g] - \delta\phi[g; f - g] \]

2. Showed Bayesian distribution estimation with functional Bregman yields mean distribution:

   \[ f^*(x) = \arg\min_{\hat{f}(x)} EF[d_{\phi}(F, \hat{f})] \equiv EF[F] \]

3. Demonstrated Bayesian distribution estimation on uniform.


5. Proposed Bayesian estimate neighborhood for completely lazy classifiers.

Extra slides
Fisher Information Metric (C. R. Rao ’45, Jeffreys ’46)

\[ dM = |I(a)|^{1/2} da \]

\( I(a) \) is the Fisher information matrix.

For the 1-d manifold \( M \) formed by the set \( \mathcal{U} \),

\[ I(a) = E_X \left[ \left( \frac{d \log 1/a}{da} \right)^2 \right] \]

\[ = \int_{x=0}^{a} \frac{1}{a^2} \frac{1}{a} dx = \frac{1}{a^2} \]

\[ \uparrow \uparrow \]

\[ g(x)p(x) \]

\[ \rightarrow dM = \frac{1}{a} \]
Fisher Information Metric (C. R. Rao '45, Jeffreys '46)

At each point on the statistical manifold $\mathcal{U}$ define a tangent, which specifies a tangent space.

If an inner product is defined on each tangent space, the collection of inner products is a Riemannian metric:

$$< \cdot, \cdot > = \{ < \cdot, \cdot >_f \mid f \in \mathcal{U} \}$$

Together with $< \cdot, \cdot >$, $\mathcal{U}$ is a Riemannian manifold.

Riemannian metric $\rightarrow$ a natural volume element $=$ measure.
The mean minimizes average squared error

Let \( x_1, x_2, \ldots, x_N \in \mathbb{R}^n \).

\[
A^* = \arg\min_{A \in \mathbb{R}^n} \frac{1}{N} \sum_j \left( \| x_j - A \|_2 \right)^2
\]

Then,

\[
A^* = \frac{1}{N} \sum_j x_j
\]

Not true of \( l_2 \) error,

\[
C^* = \arg\min_{C \in \mathbb{R}^n} \frac{1}{N} \sum_j \| x_j - C \|_2
\]

\[C^* \neq A^*\]  

\( C^* \) minimizes length of string needed to connect to points.
Bayesian Estimation of Distributions (arXiv: Frigyik, Srivastava, Gupta)

**Ex:** Given samples \( \{2, 3, 7, 8\} \), estimate the generating uniform distribution \( U[0, a] \).

Let \( F \) be a random uniform distribution: \( U[0, a] \)

Let \( p_F \) be the likelihood of \( F \) given \( N \) data samples.

\[
\begin{align*}
    f^* &= \arg \min_{\hat{f}} E_F[d_\phi[F, \hat{f}]] \\
    &\equiv E_F[F] \\
    f^*(x) &= \int_{\max(x, X_{\text{max}})}^{\infty} \left( \frac{1}{a} \right) \left( \frac{1}{a^N} \right) \frac{da}{a^{3/2}} \\
    &= \frac{1}{a^{N-2}} \frac{da}{a^{3/2}} \int_{X_{\text{max}}}^{\infty} \frac{1}{a^N} \frac{da}{a^{3/2}}
\end{align*}
\]
BDA discriminant acts like a regularized covariance estimate

\[ E_{N_h}[N_h] = \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)\left(1 + \frac{n_h}{n_h+1} Z_h^T D_h^{-1} Z_h\right) - \frac{n_h+q+1}{2}}{\pi^{\frac{d}{2}} \Gamma\left(\frac{n_h+q-d+1}{2}\right)|(\frac{n_h+1}{n_h})D_h|^\frac{1}{2}} \]

Approximate \(|Z_h^T D_h^{-1} Z_h|\) using \(1 + r \approx e^r\):

\[ E_{N_h}[N_h] \approx \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)\exp\left[-\frac{1}{2} Z_h^T \left[\frac{n_h+1}{n_h+q+1} D_h \right]^{-1} Z_h\right]}{\pi^{\frac{d}{2}} \Gamma\left(\frac{n_h+q-d+1}{2}\right)|(\frac{n_h+1}{n_h})D_h|^\frac{1}{2}} \]

\[ \tilde{\Sigma}_h = \frac{n_h+1}{n_h+q+1} \frac{D_h}{n_h} \]

\[ \approx \left(1 - \frac{q}{n_h+q+1}\right) \frac{S_h}{n_h} + \left(\frac{q}{n_h+q+1}\right) \frac{B_h}{q} \]
Figures show average distortion between each point $A$ in the space and the five black points:

$$\frac{1}{5} \sum_{j=1}^{5} d(x_j, A)$$

Bregman divergence with $\phi(x) = (\| x \|_2)^2$.

Squared Error:

$$d_\phi(x_j, A) = (\| x_j - A \|_2)^2$$

Bregman divergence with $\phi(x) = (\| x \|_2)^4$.

Results in complicated divergence function $d_\phi$.60
Functional Bregman Divergence

\( f, g : \mathbb{R}^n \to \mathbb{R} \) and \( f, g \geq 0 \), and \( f, g \in L^p(\nu) \)

\( \phi : L^p(\nu) \to \mathbb{R} \), strictly convex functional, \( \phi \in C^2 \)

\[
d_\phi(f, g) = \phi[f] - \phi[g] - \delta\phi[g; f - g]
\]

Frechet derivative of \( \phi \)
at \( g \) in the direction of \( f - g \)

Frechet derivative:

\[
\phi[g + a] - \phi[g] = \delta\phi[g; a] + \epsilon[g, a] \|a\|_{L^p(\nu)}
\]

For all \( a \in L^p(\nu) \), with \( \epsilon[g, a] \to 0 \) as \( \|a\|_{L^p(\nu)} \to 0 \).
Bayesian estimate if forced to be uniform

MER estimate solves:

\[
\arg\min_q \int_M (\|p - q\|_2)^2 P(2, 3, 7, 8|p) \, dS
\]

error if truth is \( p \)

likelihood of \( p \)

Let \( p \) be uniform from zero to \( a \):

\[
\arg\min_q \int_{a=0}^{\infty} (\|p - q\|_2)^2 P(2, 3, 7, 8|p) \left\|\frac{dp}{da}\right\|_2 \, da
\]

The MER estimate is \( q \) is uniform \( U[0,b] \):

\[
b = 2^{n+.5} k_{max}
\]

our example:

\[
b = 2^{4.5} 8
\]

\[
= 9.25
\]
Regularized Discriminant Analysis (RDA)  
(Friedman 1989)

\[
\hat{\Sigma}_h(\lambda, \gamma) = (1 - \gamma)\hat{\Sigma}_h(\lambda) + \frac{\gamma}{d}\text{trace}(\hat{\Sigma}_h(\lambda))I
\]
controls shrinkage towards a multiple of the identity

\[
\hat{\Sigma}_h(\lambda) = \frac{(1-\lambda)S_h + \lambda S}{(1-\lambda)n_h + \lambda n}
\]
controls degree of shrinkage of the class covariance matrix towards the pooled

\[
\gamma
\]

QDA  
LDA

nearest-means (identity cov)
Proposed Prior

\[ p(N_h) = p(\mu_h) p(\Sigma_h) = \gamma_0 \exp\left( -\frac{1}{2} \text{trace}(\Sigma_h^{-1} B_h) \right) \left| \Sigma_h \right|^{-q/2} \]  
\text{ (inverted Wishart)}

Differentiate \( \log p(N_h) \) with respect to \( \Sigma_h \) to solve for \( \Sigma_h,\max \):

\[
\frac{1}{2} \frac{\partial}{\partial \Sigma_h} \text{trace}(\Sigma_h^{-1} B_h) + \frac{q}{2} \frac{\partial}{\partial \Sigma_h} \log |\Sigma_h| = 0
\]

\[-\Sigma_h,\max^{-1} B_h \Sigma_h,\max^{-1} + q \Sigma_h,\max^{-1} = 0\]

\[\Sigma_h,\max = \frac{B_h}{q}\]
% Misclassification error results on UCI and Statlog benchmark datasets

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Local BDA (B = I)</th>
<th>Local BDA (B = \text{Trace})</th>
<th>Local BDA (B = \text{Diag})</th>
<th>Local Nearest Means</th>
<th>k-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glass</td>
<td>23.44</td>
<td>23.78</td>
<td>27.67</td>
<td>26.17</td>
<td>26.56</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>21.78</td>
<td>25.41</td>
<td>20.48</td>
<td>32.63</td>
<td>33.30</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>5.53</td>
<td>9.74</td>
<td>35.29</td>
<td>9.50</td>
<td>13.35</td>
</tr>
<tr>
<td>Iris</td>
<td>1.73</td>
<td>1.53</td>
<td>1.93</td>
<td>2.73</td>
<td>2.60</td>
</tr>
<tr>
<td>Letter Recognition</td>
<td>3.03</td>
<td>3.05</td>
<td>3.18</td>
<td>4.38</td>
<td>4.73</td>
</tr>
<tr>
<td>Pen Digits</td>
<td>2.17</td>
<td>2.17</td>
<td>2.26</td>
<td>2.12</td>
<td>2.66</td>
</tr>
<tr>
<td>Pima</td>
<td>27.41</td>
<td>26.68</td>
<td>26.25</td>
<td>26.80</td>
<td>26.36</td>
</tr>
<tr>
<td>Sonar</td>
<td>12.40</td>
<td>12.45</td>
<td>10.10</td>
<td>15.60</td>
<td>15.55</td>
</tr>
<tr>
<td>Thyroid</td>
<td>3.19</td>
<td>3.33</td>
<td>3.76</td>
<td>3.52</td>
<td>5.09</td>
</tr>
<tr>
<td>Vowel</td>
<td>41.77</td>
<td>36.80</td>
<td>32.72</td>
<td>41.56</td>
<td>43.51</td>
</tr>
</tbody>
</table>
Apply to Nearest-Neighbor Learning
(Gupta et al. IEEE SSP ’05)

Goal: Classify $x$ based on its $k$ nearest-neighbors such that the expected misclassification cost is minimized.

Let $\theta_h$ be the (unknown) $P(\text{class } h | \text{neighbors})$.

\[
\text{Ideal: } g^* = \arg\min_g \sum_h \text{Cost}(g, h) \theta_h
\]
Apply to Nearest-Neighbor Learning
(Gupta et al. IEEE SSP ’05)

Goal: Classify $x$ based on its $k$ nearest-neighbors such that the expected misclassification cost is minimized.

Let $\theta_h$ be the (unknown) $P(\text{class } h \mid \text{neighbors}).$

\[
\text{Ideal: } g^* = \arg \min_g \sum_h \text{Cost}(g, h) \theta_h
\]

\[
\text{Standard: } g^* = \arg \min_g \sum_h \text{Cost}(g, h) \hat{\theta}_h
\]
Apply to Nearest-Neighbor Learning
(Gupta et al. IEEE SSP ’05)

Goal: Classify $x$ based on its $k$ nearest-neighbors such that the expected misclassification cost is minimized.

Let $\theta_h$ be the (unknown) $P(\text{class } h \mid \text{neighbors})$.

\[
\text{Ideal: } g^* = \arg\min_g \sum_h \text{Cost}(g, h)\theta_h
\]

\[
\text{Standard: } g^* = \arg\min_g \sum_h \text{Cost}(g, h)\hat{\theta}_h
\]

Minimize Expected Cost:
\[
g^* = \arg\min_g E_\Theta \left[ \sum_h \text{Cost}(g, h)\Theta_h \right]
\]
Apply to Nearest-Neighbor Learning
(Gupta et al. IEEE SSP ’05)

Goal: Classify \(x\) based on its \(k\) nearest-neighbors such that the expected misclassification cost is minimized.

Let \(\theta_h\) be the (unknown) \(P(\text{class } h \mid \text{neighbors})\).

\[
\text{Ideal: } g^* = \arg\min_g \sum_h Cost(g, h) \theta_h
\]

\[
\text{Standard: } g^* = \arg\min_g \sum_h Cost(g, h) \hat{\theta}_h
\]

\[
\text{Minimize Expected Cost: } g^* = \arg\min_g E_{\Theta} \left[ \sum_h Cost(g, h) \Theta_h \right]
\]

\[
\equiv \arg\min_g \sum_h Cost(g, h) E_{\Theta} [\Theta_h]
\]
Apply to Nearest-Neighbor Learning

(Gupta, Cazzanti, Srivastava, IEEE SSP ’05)

Goal: Classify $x$ based on its $k$ nearest-neighbors such that the expected misclassification cost is minimized.

Let $\theta_h$ be the (unknown) $P(\text{class } h \mid \text{neighbors})$.

\[
\text{Ideal: } g^* = \arg\min_g \sum_h \text{Cost}(g, h) \theta_h
\]

\[
\text{Standard: } g^* = \arg\min_g \sum_h \text{Cost}(g, h) \hat{\theta}_h
\]

Minimize Expected Cost: \[
\text{Minimize Expected Cost: } g^* = \arg\min_g E_{\Theta} \left[ \sum_h \text{Cost}(g, h) \Theta_h \right]
\]

\[
\equiv \arg\min_g \sum_h \text{Cost}(g, h) E_{\Theta} [\Theta_h]
\]

\[
\equiv \arg\min_g \sum_h \text{Cost}(g, h) \left( \arg\min_{\hat{\theta}} E_{\Theta} [d(\Theta_0, \hat{\theta})] \right)
\]
How much better is MER than ML?

PMF estimate with 100 training/100 test samples on 3D Kohonen simulation

shapes: Maximum likelihood
lines: BMER estimate
MER classification results

Classification with 1000 training/1000 test and 50,000 validation samples on 4D Kohonen simulation.