Boston University
Department of Electrical and Computer Engineering

ENG EC 503 (Ishwar) Learning from Data

Assignment 2

Issued: Tue 31 Jan 2017 **Due:** 5pm Wed 8 Jan 2017 in box outside PHO440
Required reading: Your notes from lectures and slides on website.

Problem 2.1 *(MMSE, MMAE, MAP, and ML decision rules)* Let $X = Y + Z$ where Y and Z are independent Gaussians with zero means and variances σ_Y^2 and σ_Z^2 respectively which are *known*. *Derive* the MMSE, MMAE, MAP, and ML decision rules for estimating Y for a given test point $X = x$. The expressions will be functions of x, σ_Y^2 and σ_Z^2. For the MMSE estimate, also *derive* an expression for the Bayes risk $E[(Y - h_{MMSE}(X))^2]$.

Problem 2.2 *(MMSE decision rule for light intensity)* The intensity Y of a light source is exponentially distributed with a *known* mean $1/\lambda$. Given $Y = y$, let $X_1, \ldots, X_n \sim$ IID Poisson(y) be n IID photon-count measurements of the light source.

(a) *Derive* the MMSE estimate of Y given $X_1 = x_1, \ldots, X_n = x_n$, as a function of $\lambda, n, x_1, \ldots, x_n$.

(b) *Determine* the limit to which the MMSE estimate converges to as $\lambda \to 0$. *Compare* with the ML estimate of Y given $X_1 = x_1, \ldots, X_n = x_n$.

Useful result: $\forall a \in (0, \infty), \int_0^\infty t^k e^{-at} dt = \frac{k!}{a^{k+1}}$.

Problem 2.3 *(Soft Thresholding decision rule)* Let $X = Y + Z$ where $Y \perp\!\!\!\perp Z$ (i.e., Y and Z are independent), $Z \sim \mathcal{N}(0, \sigma^2)$, $\sigma > 0$, and $Y \sim p(y) = 0.5\lambda \exp(-\lambda|y|)$, $\lambda > 0$ (Note: $p(y)$ is the prior pdf of Y). *Derive* $h_{\text{MAP}}(x)$, the MAP estimate of Y based on $X = x$.

Problem 2.4 *(ML learning of Categorical model parameters)* Let $x_1, \ldots, x_n$ be n IID training samples from a categorical distribution: $p(x|\theta) = \theta_x$ where $\theta = (\theta_1, \ldots, \theta_m)^\top$ is an *unknown* pmf over the first m positive integers. *Derive* the expression for $\hat{\theta}_{ML}(x_1, \ldots, x_n)$.
Useful result: For any two pmfs $p(x)$ and $q(x)$ over the same set, the scalar quantity

$$D(p\|q) := \sum_x p(x) \ln(p(x)/q(x)) \geq 0,$$

i.e., it is always nonnegative, and is equal to zero if, and only if, $p(x) = q(x)$ for all x, i.e., if the pmfs are identical. The nonnegative scalar quantity $D(p\|q)$ is called the Kullback-Liebler (KL) divergence or relative entropy of the pair of pmfs (p, q). This is an *asymmetric* measure of distance between pmfs and plays a fundamental role in Probability, Statistics, Machine Learning, and Information Theory. A related nonnegative scalar quantity is $\sum_x p(x) \ln(1/p(x))$ which is called the entropy of the pmf p.

Problem 2.5 *(ML learning of scalar Gaussian model parameters)* Let $\hat{\mu}_{ML}$ and $\widehat{\sigma^2}_{ML}$ denote the ML estimate of the mean and variance parameters of a scalar Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ based on observing n IID training samples $x_1, \ldots, x_n$.

(a) *Derive* the expressions for $\hat{\mu}_{ML}$ and $\widehat{\sigma^2}_{ML}$ in terms of the training samples $x_1, \ldots, x_n$ and n. *Tip:* using an empirical random variable can reduce clutter.

3)

$Y \perp\!\!\!\perp Z$

$X = Y + Z$

$\underbrace{\mathcal{N}(0, \sigma^2)}_{p(z)}$

$\frac{1}{?} \lambda e^{-\lambda ?}$

$$h_{map}(x) = \arg\max_y p(y|x)$$

$\vdots$

$$h_{map}(x) = -\arg\min_y \left[-\log \ldots \right]$$

$p(x|y) = p(X = x \mid Y = y)$

$= p(Y + Z = x \mid Y = y)$

$= p(y + Z = x \mid Y = y)$

$= p(Z = x - y \mid Y = y)$

$= \underbrace{p_Z(x - y)}_{\frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}}$

$\therefore h_{map}(x)$

$$= \arg\min_y \left[\frac{(x-y)^2}{2\sigma^2} + \lambda |y| \right]$$

$$= (x-y)^2 + 2\sigma^2 \lambda |y|$$

$\tau = \sigma^2 \lambda$

$$= \underbrace{(x-y)^2 + 2\overset{\sigma^2\lambda}{\tau} + 2z|y|}_{\psi(y)}$$

candidates

1) ~~$y = \pm \infty$~~

2) $y = \psi'(y)$ exists & $\psi'(y) = 0$

3) ~~$y = \psi'(y)$ does not exist~~

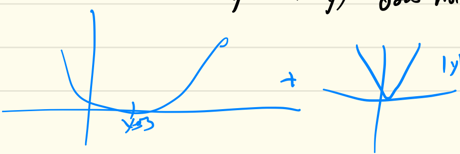
$y > 0 \Rightarrow \psi'(y) = -2(x-y) + 2\tau$

$y < 0 \Rightarrow \psi'(y) = -2(x-y) - 2\tau$

$y > 0 \Rightarrow x > \tau$

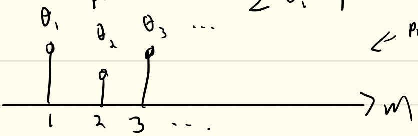
$y = x - \tau$

$x < -\tau$



4

probabilities $\sum \theta_i = 1$



pmf over m

$p(X = i) = \theta_i$

Problem 2.1 *(MMSE, MMAE, MAP, and ML decision rules)* Let $X = Y + Z$ where Y and Z are independent Gaussians with zero means and variances σ_Y^2 and σ_Z^2 respectively which are *known*. *Derive* the MMSE, MMAE, MAP, and ML decision rules for estimating Y for a given test point $X = x$. The expressions will be functions of x, σ_Y^2 and σ_Z^2. For the MMSE estimate, also *derive* an expression for the Bayes risk $E[(Y - h_{MMSE}(X))^2]$.

Feature space $\mathcal{X}$	Label space $\mathcal{Y}$	Loss function	$\ell(\mathbf{x}, y, h)$	Bayes rule name	$h_{\text{Bayes}}(\mathbf{x})$	Posterior property:
$\mathbb{R}^d$	$\mathbb{R}^m$	Squared error	$\lVert y - h(\mathbf{x})\rVert^2$	Minimum Mean Squared Error (MMSE)	$E[Y \mid \mathbf{X} = \mathbf{x}] = \int_{\mathbb{R}^m} y\, p(y\mid\mathbf{x})dy$	Mean

MMSE

$\ell(y, h(x)) = \lVert y - h(x) \rVert^2$

Examples of Bayes decision rules

- Minimum Mean Square Error (MMSE) Estimate
- Here, $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \mathbb{R}^m, \ell(\mathbf{x}, y, h) = \lVert y - h(\mathbf{x}) \rVert^2$
- Result:
  $h_{\text{Bayes}}(\mathbf{x}) = h_{\text{MMSE}}(\mathbf{x})$ = posterior mean = $E[Y \mid \mathbf{X} = \mathbf{x}]$
- *Proof*: If $g(\mathbf{x}) = E[Y|\mathbf{X}=\mathbf{x}]$ then $E[(Y - g(\mathbf{x}))|\mathbf{X}=\mathbf{x}] = 0$.
  Thus, for any h,

$$E\left[\lVert Y - h(\mathbf{x})\rVert^2 \,\middle|\, \mathbf{X} = \mathbf{x}\right] = E\left[\lVert Y - g(\mathbf{x})\rVert^2 \,\middle|\, \mathbf{X} = \mathbf{x}\right] + E\left[\lVert g(\mathbf{x}) - h(\mathbf{x}))\rVert^2 \,\middle|\, \mathbf{X} = \mathbf{x}\right]$$

$Y \perp\!\!\!\perp Z \qquad X = Y + Z$

X is observation

$MMSE = E[Y \mid X = x]$ is result

mean of posterior

$\hookrightarrow \int y\, p(y|x)\, dy$

X, Y, Z are jointly Gaussian

$\hookrightarrow$ Y & X are jointly Gaussian $\Rightarrow$ also conditionally gaussian

$y \mid X = x \sim \mathcal{N}\left(E[Y \mid X = x] \quad \mathrm{cov}(Y \mid X = x) \right)$

in slides!

prob. review slides (switch X & Y)

$$\underbrace{\mu_{y}}_{0} + \underbrace{\mathrm{Cov}(y, x)}_{E((Y+Z)(Y)) = \sigma_Y^2}\, \underbrace{(\mathrm{Cov}(x))^{-1}}_{\sigma_Y^2 + \sigma_Z^2}\, (x - \underbrace{\mu_x}_{0}) = \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_Z^2} \cdot x$$

$\underbrace{\sigma_x^2}_{\mathrm{Cov}(x, x)}$

unbiased $E[h(x)] = \mu_x$, $h(x)$ estimator

expected val of estimator is true value

2)

$$p(y \mid x_1 \ldots x_n) = \frac{p(x_1 \ldots x_n \mid y)\, p(y)}{p(x_1 \ldots x_n)}$$

$$= \frac{\prod_{j=1}^{n} p(x_j \mid y)\, p(y)}{\int_0^{\infty} Num(y)\, \partial y} = \frac{\left(\prod_{j=1}^{n} \frac{y^{x_j}}{x_j!} e^{-y}\right) \lambda e^{-\lambda y}}{\int_0^{\infty} Num(y)\, \partial y}$$

$$= y^{(x_1 + \cdots + x_n)} \Rightarrow \frac{y^{\left(\sum_{j=1}^{n} x_j\right)} e^{-ny}}{\prod_{j=1}^{n} (x_j)!} \cdot \lambda e^{-\lambda y}$$

$$= \frac{\dfrac{\cancel{\lambda} y^{\sum^{n} x_j} e^{-y(\lambda+n)}}{\cancel{\prod (x_j)!}}}{\int_0^{\infty} \dfrac{\cancel{\lambda} y^{(\,)} e^{-y(\lambda+n)}}{\cancel{\prod^{n} (x_j)!}}\, \partial y}$$

then hint: $\forall a \in (0, \infty)$

$$\int_0^{\infty} x^k e^{-at}\, \partial t = \frac{k!}{a^{k+1}}$$

$$h_{ML}(D) = \arg\max_{\theta \in \Theta} p(D \mid \theta)$$

$x \mid y$

Maximum Likelihood (ML) Decision Rule

- $h_{\mathrm{ML}}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p(\mathbf{x}|y)$
- Result: If $Y \sim \mathrm{Uniform}(\mathcal{Y})$ then the MAP rule reduces to the ML rule, i.e.,

$$h_{\mathrm{MAP}}(\mathbf{x}) = h_{\mathrm{ML}}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p(\mathbf{x}|y)$$

- *Proof*: For any $\mathbf{x}$, since Y is uniformly distributed,

$$p(y|\mathbf{x}) \propto p(\mathbf{x}|y) \cdot \underbrace{p(y)}_{\text{does not change with } y}$$

Proportional

(b) *Compute* the expected values $E[\hat{\mu}_{ML}]$ and $E[\widehat{\sigma^2}_{ML}]$ of the ML estimates with respect to the distribution on the training set. Based on your expressions, are the ML estimates of the mean and variance unbiased? If not, propose a slight modification to make them unbiased.

Problem 2.6 *(ML learning of the support of a uniform distribution)*
$X_1, \ldots, X_n \sim$ IID Uniform($[0, \theta]$) where $\theta > 0$ is a deterministic unknown parameter.

(a) Derive an ML estimate of θ given $X_1 = x_1, \ldots, X_n = x_n$ as a function of $x_1, \ldots, x_n$. Show that it is not unique. Select the one which is the smallest among all choices.

(b) Show that the ML estimate is biased. Then explain how to modify it to make it unbiased.