Recap

- Bayes classifier is optimal for minimizing risk. Risk minimization is a very good objective.

Recap

- Bayes classifier is optimal for minimizing risk. Risk minimization is a very good objective.
- Given class conditional densities we can derive the Bayes classifier for any loss function.

Recap

- Bayes classifier is optimal for minimizing risk. Risk minimization is a very good objective.
- Given class conditional densities we can derive the Bayes classifier for any loss function.
- There are other ways (other than loss function) to trade different errors. For example, NP classifier.

Recap

- Bayes classifier is optimal for minimizing risk. Risk minimization is a very good objective.
- Given class conditional densities we can derive the Bayes classifier for any loss function.
- There are other ways (other than loss function) to trade different errors. For example, NP classifier.
- ROC curve also allows for such trade-off

Receiver Operating Characteristic (ROC)

- Consider a one dimensional feature space, 2-class problem with a classifier, $h(X) = 0$ if $X < \tau$.
- Consider equal priors, Gaussian class conditional densities with equal variance, 0-1 loss. Now let us write the probability of error as a function of τ.

Receiver Operating Characteristic (ROC)

$$
\begin{aligned}
P[\text{error}] &= 0.5 \int_{-\infty}^{\tau} f_1(X)\, dX \;+\; 0.5 \int_{\tau}^{\infty} f_0(X)\, dX \\
&= 0.5\Phi\left(\frac{\tau-\mu_1}{\sigma}\right) \;+\; 0.5(1 \;-\; \Phi\left(\frac{\tau-\mu_0}{\sigma}\right))
\end{aligned}
$$

- As we vary τ we trade one kind of error with another. In Bayes classifier, the loss function determines the ‘exchange rate’.

ROC curve

- The receiver operating characteristic (ROC) curve is one way to conveniently visualize and exploit this trade off.
- For a two class classifier there are four possible outcomes of a classifcation decison – two are correct decisions and two are errors.
- Let e_i denote probability of wrongly assigning class i, $i = 0, 1$.

ROC curve

Then we have

$$
\begin{aligned}
e_0 &= P[X \leq \tau \mid X \in \textbf{c-1}] \quad \text{(a miss)} \\
e_1 &= P[X > \tau \mid X \in \textbf{c-0}] \quad \text{(false alarm)} \\
1 - e_0 &= P[X > \tau \mid X \in \textbf{c-1}] \quad \text{(correct detection)} \\
1 - e_1 &= P[X \leq \tau \mid X \in \textbf{c-0}] \quad \text{(correct rejection)}
\end{aligned}
$$

ROC curve

Then we have

$$\begin{aligned} e_0 &= P[X \leq \tau \mid X \in \textbf{c-1}] \quad \text{(a miss)} \\ e_1 &= P[X > \tau \mid X \in \textbf{c-0}] \quad \text{(false alarm)} \\ 1 - e_0 &= P[X > \tau \mid X \in \textbf{c-1}] \quad \text{(correct detection)} \\ 1 - e_1 &= P[X \leq \tau \mid X \in \textbf{c-0}] \quad \text{(correct rejection)} \end{aligned}$$

- For fixed class conditional densities, if we vary τ the point $(e_1,\ 1 - e_0)$ moves on a smooth curve in $\Re^2$.
- This is traditionally called the ROC curve. (Choice of coordinates is arbitrary)

- For any fixed τ we can estimate e_0 and e_1 from training data.

- For any fixed τ we can estimate e_0 and e_1 from training data.
- Hence, varying τ we can find ROC and decide which may be the best operating point.

- For any fixed τ we can estimate e_0 and e_1 from training data.
- Hence, varying τ we can find ROC and decide which may be the best operating point.
- This can be done for any threshold based classifier irrespective of class conditional densities.

- For any fixed τ we can estimate e_0 and e_1 from training data.
- Hence, varying τ we can find ROC and decide which may be the best operating point.
- This can be done for any threshold based classifier irrespective of class conditional densities.
- When the class conditional densities are Gaussian with equal variance, we use this procedure to estimate Bayes error also.

- From our earlier error integral we get

$$
\begin{aligned}
\frac{\tau - \mu_0}{\sigma} &= \Phi^{-1}(1 - e_1) = a, \text{ say} \\
\frac{\tau - \mu_1}{\sigma} &= \Phi^{-1}(1 - (1 - e_0)) = b, \text{ say}
\end{aligned}
$$

- From our earlier error integral we get

$$\frac{\tau - \mu_0}{\sigma} = \Phi^{-1}(1 - e_1) = a, \text{ say}$$
$$\frac{\tau - \mu_1}{\sigma} = \Phi^{-1}(1 - (1 - e_0)) = b, \text{ say}$$

- Then, $|a - b| = \frac{|\mu_1 - \mu_0|}{\sigma} = d$, the discriminability.

- From our earlier error integral we get

$$\frac{\tau - \mu_0}{\sigma} = \Phi^{-1}(1 - e_1) = a, \text{ say}$$
$$\frac{\tau - \mu_1}{\sigma} = \Phi^{-1}(1 - (1 - e_0)) = b, \text{ say}$$

- Then, $|a - b| = \frac{|\mu_1 - \mu_0|}{\sigma} = d$, the discriminability.
- Knowing $e_1, (1 - e_0)$, we can get d and hence the Bayes error. For our given τ we can also get the actuall error probability. We can tweak τ to match the Bayes error.

- We can in general use the ROC curve in multidimensional cases also. Consider, for example,

  $h(\mathbf{X}) = \text{sgn}(\mathbf{W}^t\mathbf{X} + w_0)$.

  We can use ROC to fix w_0 after learning $\mathbf{W}$.

Implementing Bayes Classifier

- We need class conditional densities and prior probabilities.

Implementing Bayes Classifier

- We need class conditional densities and prior probabilities.
- Prior probabilities can be estimated as fraction of examples from each class.

Implementing Bayes Classifier

- We need class conditional densities and prior probabilities.
- Prior probabilities can be estimated as fraction of examples from each class.
- Since examples are *iid* and the class labels of examples are known, we have some iid samples from each class conditional distribution.

Implementing Bayes Classifier

- We need class conditional densities and prior probabilities.
- Prior probabilities can be estimated as fraction of examples from each class.
- Since examples are *iid* and the class labels of examples are known, we have some iid samples from each class conditional distribution.
- The problem: Given $\{x_1, x_2, \cdots, x_n\}$ drawn *iid* according to some distribution, estimate the probability distribution / density.

Estimating densities

- Two main approaches: Parametric and non-parametric.

Estimating densities

- Two main approaches: Parametric and non-parametric.
- Parametric: We assume we have *iid* realizations of a random variable X whose distribution is known except for values of a parameter vector. We estimate the parameters of the density using the samples available.

Estimating densities

- Two main approaches: Parametric and non-parametric.
- Parametric: We assume we have *iid* realizations of a random variable X whose distribution is known except for values of a parameter vector. We estimate the parameters of the density using the samples available.
- In non-parametric approach we do not assume form of density. It is often modelled as a convex combination of some densities using the samples.

Estimating parameters of a density

- Denote the density by $f(x \mid \theta)$ where θ is a parameter vector.

Estimating parameters of a density

- Denote the density by $f(x \mid \theta)$ where θ is a parameter vector.
- For example, let $\theta = (\theta_1,\ \theta_2)$ and

$$f(x \mid \theta) = \frac{1}{2\pi\sqrt{\theta_2}} \exp\left(-\frac{(x-\theta_1)^2}{2\theta_2}\right)$$

Estimating parameters of a density

- Denote the density by $f(x \mid \theta)$ where θ is a parameter vector.
- For example, let $\theta = (\theta_1,\ \theta_2)$ and

$$f(x \mid \theta) = \frac{1}{2\pi\sqrt{\theta_2}} \exp\left(-\frac{(x-\theta_1)^2}{2\theta_2}\right)$$

$f(x|\theta)$ is normal with mean and variance constituting the parameter vector.

Estimating parameters of a density

- Denote the density by $f(x \mid \theta)$ where θ is a parameter vector.
- For example, let $\theta = (\theta_1,\ \theta_2)$ and

$$f(x \mid \theta) = \frac{1}{2\pi\sqrt{\theta_2}} \exp\left(-\frac{(x-\theta_1)^2}{2\theta_2}\right)$$

  $f(x|\theta)$ is normal with mean and variance constituting the parameter vector.
- Now estimation of density is same as estimation of a parameter vector.

Notation

- Let X denote a random variable with density $f(x \mid \theta)$. (Use same notation even when X is a random vector)

Notation

- Let X denote a random variable with density $f(x \mid \theta)$. (Use same notation even when X is a random vector)
- A (*iid*) sample of size n consists of n *iid* realizations of X.

Notation

- Let X denote a random variable with density $f(x \mid \theta)$. (Use same notation even when X is a random vector)
- A (*iid*) sample of size n consists of n *iid* realizations of X.
- $\mathbf{x} = (x_1, \cdots, x_n)^T$ – the sample or the data. We sometimes use $\mathcal{D}$ to denote the data.

Notation

- Let X denote a random variable with density $f(x \mid \theta)$. (Use same notation even when X is a random vector)
- A (*iid*) sample of size n consists of n *iid* realizations of X.
- $\mathbf{x} = (x_1, \cdots, x_n)^T$ – the sample or the data. We sometimes use $\mathcal{D}$ to denote the data.
- It can be thought of as a realization of $(X_1, \cdots, X_n)^T$ where X_i are *iid* with density $f(x \mid \theta)$.

- A *statistic* is a function of data, e.g., $g(x_1, \cdots, x_n)$.

- A *statistic* is a function of data, e.g., $g(x_1, \cdots, x_n)$.
- An estimator is such a statistic. $\hat{\theta}(x_1, \cdots, x_n)$.

- A *statistic* is a function of data, e.g., $g(x_1, \cdots, x_n)$.
- An estimator is such a statistic. $\hat{\theta}(x_1, \cdots, x_n)$.
- When we need to remember the sample size, we write $\hat{\theta}_n$

- A *statistic* is a function of data, e.g., $g(x_1, \cdots, x_n)$.
- An estimator is such a statistic. $\hat{\theta}(x_1, \cdots, x_n)$.
- When we need to remember the sample size, we write $\hat{\theta}_n$
- For example,

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

the well-known sample mean.

- There can be different estimators that are intuitively reasonable.

- There can be different estimators that are intuitively reasonable.
- Let X be Poisson with parameter λ. Then sample mean as well as sample variance seem to be reasonable estimators for λ.

- There can be different estimators that are intuitively reasonable.
- Let X be Poisson with parameter λ. Then sample mean as well as sample variance seem to be reasonable estimators for λ.
- Let X be normal with mean μ and variance unity. Both sample mean and sample median seem good choices.

- There can be different estimators that are intuitively reasonable.
- Let X be Poisson with parameter λ. Then sample mean as well as sample variance seem to be reasonable estimators for λ.
- Let X be normal with mean μ and variance unity. Both sample mean and sample median seem good choices.
- How does one choose estimators

- We need ‘good’ estimators.

- We need ‘good’ estimators.
- We need some criteria for ‘goodness’. Also, methods to obtain such estimators.

- We need ‘good’ estimators.
- We need some criteria for ‘goodness’. Also, methods to obtain such estimators.
- In this course, we will consider two methods: Maximum likelihood and Bayesian estimators.

- We need ‘good’ estimators.
- We need some criteria for ‘goodness’. Also, methods to obtain such estimators.
- In this course, we will consider two methods: Maximum likelihood and Bayesian estimators.
- To begin with, a simple introduction to some general issues in estimation.

- An estimator, $\hat{\theta}$ of a parameter (vector) θ is said to be **unbiased** if $E[\hat{\theta}] = \theta$.

- An estimator, $\hat{\theta}$ of a parameter (vector) θ is said to be **unbiased** if $E[\hat{\theta}] = \theta$.
- The $\hat{\theta}$ is a function of data. Hence the expectation is with respect to the joint density of $(X_1, \cdots X_n)$, the *iid* random variables.

- An estimator, $\hat{\theta}$ of a parameter (vector) θ is said to be **unbiased** if $E[\hat{\theta}] = \theta$.
- The $\hat{\theta}$ is a function of data. Hence the expectation is with respect to the joint density of $(X_1, \cdots X_n)$, the *iid* random variables.
- Since $X_i \sim f(x \mid \theta)$, the expectation above needs value of θ. So, we write

$$E_\theta[\hat{\theta}] = \theta$$

- An unbiased estimator, $\hat{\theta}$ satisfies

$$E_\theta[\hat{\theta}] = \theta$$

- $\hat{\theta}$ is an unbiased estimator, if for every density in the class of densities we are interested in (i.e., every value of the parameter in the parameter space), expected value of the estimator is the true parameter value.

- Let $f(x \mid \theta)$ be normal with mean θ and variance unity. Let $\hat{\theta}_n = (1/n) \sum_i x_i$

- Let $f(x \mid \theta)$ be normal with mean θ and variance unity. Let $\hat{\theta}_n = (1/n)\sum_i x_i$
- Then $E[\hat{\theta}_n] = \theta$ for all n because $EX_i = \theta$.

- Let $f(x \mid \theta)$ be normal with mean θ and variance unity. Let $\hat{\theta}_n = (1/n)\sum_i x_i$
- Then $E[\hat{\theta}_n] = \theta$ for all n because $EX_i = \theta$.
- Sample mean is an unbiased estimator of actual mean.

- Let $f(x \mid \theta)$ be normal with mean θ and variance unity. Let $\hat{\theta}_n = (1/n) \sum_i x_i$
- Then $E[\hat{\theta}_n] = \theta$ for all n because $EX_i = \theta$.
- Sample mean is an unbiased estimator of actual mean.
- Let $\hat{\theta}'(x_1, \cdots, x_n) = 0.5(x_1 + x_2)$.
- This is also an unbiased estimator.

- Let $f(x \mid \theta)$ be normal with mean θ and variance unity. Let $\hat{\theta}_n = (1/n) \sum_i x_i$
- Then $E[\hat{\theta}_n] = \theta$ for all n because $EX_i = \theta$.
- Sample mean is an unbiased estimator of actual mean.
- Let $\hat{\theta}'(x_1, \cdots, x_n) = 0.5(x_1 + x_2)$.
- This is also an unbiased estimator.
- So is $\hat{\theta}'' = x_1$.

- Let $f(x \mid \theta)$ be normal with mean θ and variance unity. Let $\hat{\theta}_n = (1/n)\sum_i x_i$
- Then $E[\hat{\theta}_n] = \theta$ for all n because $EX_i = \theta$.
- Sample mean is an unbiased estimator of actual mean.
- Let $\hat{\theta}'(x_1, \cdots, x_n) = 0.5(x_1 + x_2)$.
- This is also an unbiased estimator.
- So is $\hat{\theta}'' = x_1$.
- Unbiasedness alone is not enough

- One possibility: We can say $\hat{\theta}$ is better than $\hat{\theta}'$ if, $\forall \theta$,

$$P_\theta[-a \leq (\hat{\theta} - \theta) \leq b] \geq P_\theta[-a \leq (\hat{\theta}' - \theta) \leq b] \ \forall a, b > 0$$

(for any fixed sample size)

- One possibility: We can say $\hat{\theta}$ is better than $\hat{\theta}'$ if $\forall\theta$,

  $$P_\theta[-a \leq (\hat{\theta}-\theta) \leq b] \geq P_\theta[-a \leq (\hat{\theta}'-\theta) \leq b] \ \forall a, b > 0$$

  (for any fixed sample size)
- Difficult to get such estimators.

- A weaker method is: $\hat{\theta}$ is better than $\hat{\theta}'$ if

$$E_\theta[(\hat{\theta} - \theta)^2] \leq E_\theta[(\hat{\theta}' - \theta)^2] \; \forall \theta$$

- A weaker method is: $\hat{\theta}$ is better than $\hat{\theta}'$ if

$$E_\theta[(\hat{\theta} - \theta)^2] \leq E_\theta[(\hat{\theta}' - \theta)^2] \; \forall \theta$$

- The mean square error of an estimator is defined by

$$\text{MSE}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2]$$

- Lemma:

$$\text{MSE}_\theta(\hat{\theta}) = V_\theta(\hat{\theta}) + [B_\theta(\hat{\theta})]^2$$

where $V_\theta(\hat{\theta})$ is the variance given by

$$V_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$$

and $B_\theta(\hat{\theta})$ is the bias given by

$$B_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta$$

- Lemma:

$$\text{MSE}_\theta(\hat{\theta}) = V_\theta(\hat{\theta}) + [B_\theta(\hat{\theta})]^2$$

where $V_\theta(\hat{\theta})$ is the variance given by

$$V_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$$

and $B_\theta(\hat{\theta})$ is the bias given by

$$B_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta$$

- For unbiased estimators the variance is the mean square error (because bias is zero).

- Proof:

$$\text{MSE}(\hat{\theta}) \;=\; E[(\hat{\theta} - \theta)^2]$$

- Proof:

$$
\begin{aligned}
\mathsf{MSE}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E[\{(\hat{\theta} - E[\hat{\theta}]) \;+\; (E[\hat{\theta}] - \theta)\}^2]
\end{aligned}
$$

- Proof:

$$
\begin{aligned}
\text{MSE}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E[\{(\hat{\theta} - E[\hat{\theta}]) \;+\; (E[\hat{\theta}] - \theta)\}^2] \\
&= E[(\hat{\theta} - E[\hat{\theta}])^2] \;+\; (E[\hat{\theta}] - \theta)^2 \;+ \\
&\qquad\qquad 2E\left[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)\right]
\end{aligned}
$$

- Proof:

$$
\begin{aligned}
\text{MSE}(\hat{\theta}) &= E[(\hat{\theta}-\theta)^2] \\
&= E[\{(\hat{\theta}-E[\hat{\theta}]) + (E[\hat{\theta}]-\theta)\}^2] \\
&= E[(\hat{\theta}-E[\hat{\theta}])^2] + (E[\hat{\theta}]-\theta)^2 + \\
&\qquad 2E\left[(\hat{\theta}-E[\hat{\theta}])(E[\hat{\theta}]-\theta)\right] \\
&= V(\hat{\theta}) + [B(\hat{\theta})]^2 + 2(E[\hat{\theta}]-\theta)E[(\hat{\theta}-E[\hat{\theta}])
\end{aligned}
$$

- Proof:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E[(\hat{\theta}-\theta)^2] \\
&= E[\{(\hat{\theta}-E[\hat{\theta}]) \; + \; (E[\hat{\theta}]-\theta)\}^2] \\
&= E[(\hat{\theta}-E[\hat{\theta}])^2] \; + \; (E[\hat{\theta}]-\theta)^2 \; + \\
&\qquad 2E\left[(\hat{\theta}-E[\hat{\theta}])(E[\hat{\theta}]-\theta)\right] \\
&= V(\hat{\theta}) \; + \; [B(\hat{\theta})]^2 \; + \; 2(E[\hat{\theta}]-\theta)E[(\hat{\theta}-E[\hat{\theta}]) \\
&= V(\hat{\theta}) \; + \; [B(\hat{\theta})]^2
\end{aligned}$$

- For unbiased estimators, low variance implies low MSE.

- For unbiased estimators, low variance implies low MSE.
- Earlier example: When $\hat{\theta}$ is the sample mean,

$$V_\theta(\hat{\theta}_n) = \frac{\sigma^2}{n}$$

- For unbiased estimators, low variance implies low MSE.
- Earlier example: When $\hat{\theta}$ is the sample mean,

$$V_\theta(\hat{\theta}_n) = \frac{\sigma^2}{n}$$

For $\hat{\theta}'_n = 0.5(x_1 + x_2)$,

$$V_\theta(\hat{\theta}'_n) = \frac{\sigma^2}{2}$$

- For unbiased estimators, low variance implies low MSE.
- Earlier example: When $\hat{\theta}$ is the sample mean,

$$V_\theta(\hat{\theta}_n) = \frac{\sigma^2}{n}$$

For $\hat{\theta}'_n = 0.5(x_1 + x_2)$,

$$V_\theta(\hat{\theta}'_n) = \frac{\sigma^2}{2}$$

- Hence $\hat{\theta}$ is better than $\hat{\theta}'$

- So, unbiased estimators with low mean square error are good.

- So, unbiased estimators with low mean square error are good.
- For a given family of density functions, $\hat{\theta}$ is said to be **uniformly minimum variance unbiased estimator (UMVUE)** if

1. $\hat{\theta}$ is unbiased, and

- So, unbiased estimators with low mean square error are good.
- For a given family of density functions, $\hat{\theta}$ is said to be **uniformly minimum variance unbiased estimator (UMVUE)** if
  1. $\hat{\theta}$ is unbiased, and
  2. $\text{MSE}_\theta(\hat{\theta}_n) \leq \text{MSE}_\theta(\hat{\theta}'_n) \; \forall n, \theta$,
     and forall $\hat{\theta}'$ that are unbiased estimators for θ.

- So, unbiased estimators with low mean square error are good.
- For a given family of density functions, $\hat{\theta}$ is said to be **uniformly minimum variance unbiased estimator (UMVUE)** if
  1. $\hat{\theta}$ is unbiased, and
  2. MSE$_\theta(\hat{\theta}_n) \leq$ MSE$_\theta(\hat{\theta}'_n)$ $\forall n, \theta$,

     and forall $\hat{\theta}'$ that are unbiased estimators for θ.
- If we can get an UMVUE, then it is the 'best' estimator.
- In many cases, it is difficult to get UMVUE.

- So far, we are looking at figures of merit of estimators at (all) fixed sample sizes.

- So far, we are looking at figures of merit of estimators at (all) fixed sample sizes.
- We can also think of asymptotic properties.

- So far, we are looking at figures of merit of estimators at (all) fixed sample sizes.
- We can also think of asymptotic properties.
- An estimator $\hat{\theta}$ is said to be **consistent** for θ if

$$\hat{\theta}_n \xrightarrow{P} \theta \ \ \forall \theta$$

- So far, we are looking at figures of merit of estimators at (all) fixed sample sizes.
- We can also think of asymptotic properties.
- An estimator $\hat{\theta}$ is said to be **consistent** for θ if

$$\hat{\theta}_n \xrightarrow{P} \theta \;\; \forall \theta$$

- For example, the sample mean is a consistent estimator of population mean (expectation of the random variable)
(Law of large numbers)

- A consistent estimator need not be unbiased.

- A consistent estimator need not be unbiased.
- Let θ be the mean and let

$$\hat{\theta}_n = \frac{1}{n+1} \sum_{i=1}^{n} x_i$$

- A consistent estimator need not be unbiased.
- Let θ be the mean and let

$$\hat{\theta}_n = \frac{1}{n+1} \sum_{i=1}^{n} x_i$$

- This is not an unbiased estimator.

- A consistent estimator need not be unbiased.
- Let θ be the mean and let

$$\hat{\theta}_n = \frac{1}{n+1} \sum_{i=1}^{n} x_i$$

- This is not an unbiased estimator.
- But we have the following

$$E[(\hat{\theta}_n - \theta)^2] \;=\; E\left[\left(\frac{1}{n+1}\sum_{i=1}^{n}(x_i - \theta) \;-\; \frac{1}{n+1}\theta\right)^2\right]$$

$$
\begin{aligned}
E[(\hat{\theta}_n - \theta)^2] &= E\left[\left(\frac{1}{n+1}\sum_{i=1}^{n}(x_i - \theta) \; - \; \frac{1}{n+1}\theta\right)^2\right] \\
&= \frac{1}{(n+1)^2} n\sigma^2 \; + \; \frac{1}{(n+1)^2}\theta^2 \; - \\
&\qquad \frac{2\theta}{(n+1)^2} E[\sum (x_i - \theta)]
\end{aligned}
$$

$$
\begin{aligned}
E[(\hat{\theta}_n - \theta)^2] &= E\left[\left(\frac{1}{n+1}\sum_{i=1}^{n}(x_i - \theta) \; - \; \frac{1}{n+1}\theta\right)^2\right] \\
&= \frac{1}{(n+1)^2} n\sigma^2 \; + \; \frac{1}{(n+1)^2}\theta^2 - \\
&\qquad \frac{2\theta}{(n+1)^2} E[\sum (x_i - \theta)] \\
&= \frac{n}{(n+1)^2}\sigma^2 + \frac{1}{(n+1)^2}\theta^2
\end{aligned}
$$

- Thus, $E[(\hat{\theta}_n - \theta)^2] \rightarrow 0$ as $n \rightarrow \infty$.
- Hence, $\hat{\theta}$ is consistent (though it is biased).

- Maximum Likelihood (ML) estimation is a general procedure for obtaining consistent estimators.
- It is a parametric method.
- We estimate parameters of a density based on *iid* samples.
- For most densities, ML estimates are consistent.

Maximum likelihood estimation

- Let $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ be the samples.
- Likelihood function is defined by

$$L(\mathbf{x}, \theta) = \prod_{j=1}^{n} f(x_j|\theta)$$

Maximum likelihood estimation

- Let $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ be the samples.
- Likelihood function is defined by

$$L(\mathbf{x}, \theta) = \prod_{j=1}^{n} f(x_j|\theta)$$

- If samples are from a discrete random variable, f is taken to be the mass function. If samples are from a continuous random variable, then f is the density function.

Maximum likelihood estimation

- We essentially look at the likelihood function as a function of θ with the x_j being known values (as given by data).

Maximum likelihood estimation

- We essentially look at the likelihood function as a function of θ with the x_j being known values (as given by data).
- To emphasize this we write it as $L(\theta, \mathbf{x})$ or $L(\theta \mid \mathbf{x})$ or $L(\theta \mid \mathcal{D})$.
  Recall that we denote the data samples by $\mathcal{D}$ also.

Maximum likelihood estimation contd..

- The maximum likelihood (ML) estimate of θ is the value that (globally) maximizes the likelihood function.

Maximum likelihood estimation contd..

- The maximum likelihood (ML) estimate of θ is the value that (globally) maximizes the likelihood function.
- θ^* is the MLE for θ if

$$L(\theta^* \mid \mathbf{x}) \geq L(\theta \mid \mathbf{x}) \;\; \forall \theta$$

Maximum likelihood estimation contd..

- The maximum likelihood (ML) estimate of θ is the value that (globally) maximizes the likelihood function.
- θ^* is the MLE for θ if

$$L(\theta^* \mid \mathbf{x}) \geq L(\theta \mid \mathbf{x}) \ \ \forall \theta$$

- Finding MLE is an optimization problem.

- For convenience in optimization we often take the log likelihood given by

$$l(\theta \mid \mathbf{x}) = \log L(\theta \mid \mathbf{x}) = \sum_{j=1}^{n} \log f(x_j|\theta)$$

- For convenience in optimization we often take the log likelihood given by

$$l(\theta \mid \mathbf{x}) = \log L(\theta \mid \mathbf{x}) = \sum_{j=1}^{n} \log f(x_j|\theta)$$

- Now the ML estimate would be maximizer of the log likelihood.

- For convenience in optimization we often take the log likelihood given by

$$l(\theta \mid \mathbf{x}) = \log L(\theta \mid \mathbf{x}) = \sum_{j=1}^{n} \log f(x_j|\theta)$$

- Now the ML estimate would be maximizer of the log likelihood.
- For many densities we can analytically solve for the maximizer.

- For convenience in optimization we often take the log likelihood given by

$$l(\theta \mid \mathbf{x}) = \log L(\theta \mid \mathbf{x}) = \sum_{j=1}^{n} \log f(x_j|\theta)$$

- Now the ML estimate would be maximizer of the log likelihood.
- For many densities we can analytically solve for the maximizer.
- In general we can use numerical optimization techniques.

Example

- Consider one dimensional case.
  Let $f(x|\theta) \sim \mathcal{N}(\mu, \sigma^2)$ with $\theta_1 = \mu$ and $\theta_2 = \sigma$.

Example

- Consider one dimensional case.
  Let $f(x|\theta) \sim \mathcal{N}(\mu, \sigma^2)$ with $\theta_1 = \mu$ and $\theta_2 = \sigma$.

$$f(x|\theta) = \frac{1}{\theta_2\sqrt{2\pi}} exp\left(-\frac{(x-\theta_1)^2}{2\theta_2^2}\right)$$

Example

- Consider one dimensional case.
  Let $f(x|\theta) \sim \mathcal{N}(\mu, \sigma^2)$ with $\theta_1 = \mu$ and $\theta_2 = \sigma$.

$$f(x|\theta) = \frac{1}{\theta_2\sqrt{2\pi}} exp\left(-\frac{(x-\theta_1)^2}{2\theta_2^2}\right)$$

- Now the likelihood is given by

$$L(\theta \mid \mathbf{x}) = \prod_{j=1}^{n} \frac{1}{\theta_2\sqrt{2\pi}} exp\left(-\frac{(x_j-\theta_1)^2}{2\theta_2^2}\right)$$

- Hence log likelihood would be

$$l(\theta \mid \mathbf{x}) \;=\; \sum_{j=1}^{n} \left[-\log(\theta_2) - 0.5\log(2\pi) - \frac{(x_j - \theta_1)^2}{2\theta_2^2} \right]$$

Example

- Hence log likelihood would be

$$
\begin{aligned}
l(\theta \mid \mathbf{x}) &= \sum_{j=1}^{n} \left[-\log(\theta_2) - 0.5\log(2\pi) - \frac{(x_j - \theta_1)^2}{2\theta_2^2} \right] \\
&= -n\log(\theta_2) - 0.5n\log(2\pi) - \sum_{j=1}^{n} \frac{(x_j - \theta_1)^2}{2\theta_2^2}
\end{aligned}
$$

Example

- Hence log likelihood would be

$$
\begin{aligned}
l(\theta \mid \mathbf{x}) &= \sum_{j=1}^{n} \left[-\log(\theta_2) - 0.5\log(2\pi) - \frac{(x_j - \theta_1)^2}{2\theta_2^2} \right] \\
&= -n\log(\theta_2) - 0.5n\log(2\pi) - \sum_{j=1}^{n} \frac{(x_j - \theta_1)^2}{2\theta_2^2}
\end{aligned}
$$

- To maximize log likelihood we equate the partial derivatives to zero.

- This gives

$$\frac{\partial l}{\partial \theta_1} = \sum_{j=1}^{n} (x_j - \theta_1) = 0$$

$$\frac{\partial l}{\partial \theta_2} = -\frac{n}{\theta_2} + \frac{1}{\theta_2^3} \sum_{j=1}^{n} (x_j - \theta_1)^2 = 0$$

- Solving these, we get

$$\hat{\theta}_1 = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{j=1}^{n} (x_j - \hat{\theta}_1)^2$$

- Solving these, we get

$$\hat{\theta}_1 = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{j=1}^{n} (x_j - \hat{\theta}_1)^2$$

- These are the ML estimates of mean and variance of a normal density

- Solving these, we get

$$\hat{\theta}_1 = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{j=1}^{n} (x_j - \hat{\theta}_1)^2$$

- These are the ML estimates of mean and variance of a normal density
- ML estimate of variance is **not** unbiased.

Example: discrete case

- Let X have Bernoulli distribution. That is X takes values 0 and 1 with probability $(1-p)$ and p respectively.

Example: discrete case

- Let X have Bernoulli distribution. That is X takes values 0 and 1 with probability $(1-p)$ and p respectively.
- Then, $f(x|p) = p^x(1-p)^{1-x},\ x \in \{0,\ 1\}$

Example: discrete case

- Let X have Bernoulli distribution. That is X takes values 0 and 1 with probability $(1-p)$ and p respectively.
- Then, $f(x|p) = p^x(1-p)^{1-x},\ x \in \{0,\ 1\}$
- The mass function has only one parameter, namely, p.

Example: discrete case

- Let X have Bernoulli distribution. That is X takes values 0 and 1 with probability $(1-p)$ and p respectively.
- Then, $f(x|p) = p^x(1-p)^{1-x},\ x \in \{0,\ 1\}$
- The mass function has only one parameter, namely, p.
- Note that we must have $0 \leq p \leq 1$.

- The likelihood function is

$$L(p \mid \mathbf{x}) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^{n\bar{x}}(1-p)^{n-n\bar{x}}$$

where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ is the sample mean.

- The likelihood function is

$$L(p \mid \mathbf{x}) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^{n\bar{x}}(1-p)^{n-n\bar{x}}$$

where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ is the sample mean.

- The loglikelihood is given by

$$l(p \mid \mathbf{x}) = n\bar{x}\log p + n(1-\bar{x})\log(1-p)$$

- Differentiating the log likelihood with respect to p and equating to zero we get

$$\frac{n\bar{x}}{p} = \frac{n(1-\bar{x})}{1-p}$$

- Differentiating the log likelihhod with respect to p and equating to zero we get

$$\frac{n\bar{x}}{p} = \frac{n(1-\bar{x})}{1-p}$$

which implies

$$p = \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

- Differentiating the log likelihood with respect to p and equating to zero we get

$$\frac{n\bar{x}}{p} = \frac{n(1-\bar{x})}{1-p}$$

which implies

$$p = \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

- This is the ML estimate of the parameter p of a Bernoulli random variable.
- Sample mean is the ML estimator.

To Summarize

- To implement Bayes classifier, we need to estimate densities.

To Summarize

- To implement Bayes classifier, we need to estimate densities.
- Parametric methods assume that form of density is known.

To Summarize

- To implement Bayes classifier, we need to estimate densities.
- Parametric methods assume that form of density is known.
- Estimate (for a parameter) is a function of (*iid*) data

To Summarize

- To implement Bayes classifier, we need to estimate densities.
- Parametric methods assume that form of density is known.
- Estimate (for a parameter) is a function of (*iid*) data
- An estimate is unbiased if its expectation is the true value.

To Summarize

- To implement Bayes classifier, we need to estimate densities.
- Parametric methods assume that form of density is known.
- Estimate (for a parameter) is a function of (*iid*) data
- An estimate is unbiased if its expectation is the true value.
- The MSE of an unbiased estimator is its variance.

To Summarize

- To implement Bayes classifier, we need to estimate densities.
- Parametric methods assume that form of density is known.
- Estimate (for a parameter) is a function of (*iid*) data
- An estimate is unbiased if its expectation is the true value.
- The MSE of an unbiased estimator is its variance.
- UMVUE is a good estimate to have

- Consistent estimators converge to the true value in probability as sample size goes to infinity

- Consistent estimators converge to the true value in probability as sample size goes to infinity
- Maximum likelihood estimation is a general procedure that can find consistent estimators.

- Consistent estimators converge to the true value in probability as sample size goes to infinity
- Maximum likelihood estimation is a general procedure that can find consistent estimators.
- MLE is the maximizer of the likelihood function.

- Consistent estimators converge to the true value in probability as sample size goes to infinity
- Maximum likelihood estimation is a general procedure that can find consistent estimators.
- MLE is the maximizer of the likelihood function.
- Often, one maximizes loglikelihood

- Consistent estimators converge to the true value in probability as sample size goes to infinity
- Maximum likelihood estimation is a general procedure that can find consistent estimators.
- MLE is the maximizer of the likelihood function.
- Often, one maximizes loglikelihood
- For many standard densities we can obtain MLE analytically.