EK381 Exam 3 Formula Sheet

6. Detection

- Two hypotheses H_0 and H_1. Observe a random variable Y. Decide if H_0 or H_1 occurred based only on Y using a decision rule $D(y)$.

Discrete Case	Continuous Case
$P_{Y\|H_0}(y)$ if H_0 occurs	$f_{Y\|H_0}(y)$ if H_0 occurs
$P_{Y\|H_1}(y)$ if H_1 occurs	$f_{Y\|H_1}(y)$ if H_1 occurs

- Decision Regions: $A_0 = \{y \in R_Y : D(y) = 0\}$ $\quad A_1 = \{y \in R_Y : D(y) = 1\}$
- Probability of False Alarm: $P_{\text{FA}} = \mathbb{P}[Y \in A_1 | H_0]$
- Probability of Missed Detection: $P_{\text{MD}} = \mathbb{P}[Y \in A_0 | H_1]$
- Goal is to minimize the probability of error:

$$P_e = \mathbb{P}[\{\text{error}\}] = P_{\text{FA}}\,\mathbb{P}[H_0] + P_{\text{MD}}\,\mathbb{P}[H_1]$$

- Likelihood Ratio: $L(y) = \dfrac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)}$.
- Log-Likelihood Ratio: $\ln\big(L(y)\big) = \ln\left(\dfrac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)}\right)$.
- For vector observations $\underline{Y}$, we simply replace all occurrences of Y with $\underline{Y}$. For example, $P_{Y|H_0}(y)$ becomes $P_{\underline{Y}|H_0}(\underline{y})$ and $P_{Y|H_1}(y)$ becomes $P_{\underline{Y}|H_1}(\underline{y})$.

Maximum Likelihood (ML) Rule

- Intuition: Choose hypothesis that best explains Y.
- In terms of the conditional PMFs for the discrete case,

$$D^{\text{ML}}(y) = \begin{cases} 1, & P_{Y|H_1}(y) \geq P_{Y|H_0}(y), \\ 0, & P_{Y|H_1}(y) < P_{Y|H_0}(y). \end{cases}$$

- In terms of the conditional PDFs for the continuous case,

$$D^{\text{ML}}(y) = \begin{cases} 1, & f_{Y|H_1}(y) \geq f_{Y|H_0}(y), \\ 0, & f_{Y|H_1}(y) < f_{Y|H_0}(y). \end{cases}$$

- In terms of the likelihood or log-likelihood ratio,

$$D^{\text{ML}}(y) = \begin{cases} 1, & L(y) \geq 1, \\ 0, & L(y) < 1. \end{cases} = \begin{cases} 1, & \ln(L(y)) \geq 0, \\ 0, & \ln(L(y)) < 0. \end{cases}$$

Maximum a Posteriori (MAP) Rule

- Intuition: Choose the most likely hypothesis.
- In terms of the conditional PMFs for the discrete case,

$$D^{\text{MAP}}(y) = \begin{cases} 1, & P_{Y|H_1}(y)\,\mathbb{P}[H_1] \geq P_{Y|H_0}(y)\,\mathbb{P}[H_0], \\ 0, & P_{Y|H_1}(y)\,\mathbb{P}[H_1] < P_{Y|H_0}(y)\,\mathbb{P}[H_0]. \end{cases}$$

- In terms of the conditional PDFs for the continuous case,

$$D^{\text{MAP}}(y) = \begin{cases} 1, & f_{Y|H_1}(y)\,\mathbb{P}[H_1] \geq f_{Y|H_0}(y)\,\mathbb{P}[H_0], \\ 0, & f_{Y|H_1}(y)\,\mathbb{P}[H_1] < f_{Y|H_0}(y)\,\mathbb{P}[H_0]. \end{cases}$$

- In terms of the likelihood or log-likelihood ratio,

$$D^{\text{MAP}}(y) = \begin{cases} 1, & L(y) \geq \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]}, \\ 0, & L(y) < \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]}. \end{cases} = \begin{cases} 1, & \ln(L(y)) \geq \ln\left(\frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]}\right), \\ 0, & \ln(L(y)) < \ln\left(\frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]}\right). \end{cases}$$

- This is the optimal decision rule in terms of minimizing the probability of error. However, it requires knowledge of $\mathbb{P}[H_0]$ and $\mathbb{P}[H_1]$ to implement the decision rule.

7. Estimation

- We observe a random variable Y and want to estimate an unobserved random variable X using an estimator $\hat{x}(Y)$.
- Goal: Minimize the mean-squared error:

$$\mathsf{MSE} = \mathbb{E}\big[(X - \hat{x}(Y))^2\big]$$

MMSE Estimator

- The minimum mean-squared error (MMSE) estimator is

$$\hat{x}_{\text{MMSE}}(y) = \mathbb{E}\big[X | Y = y\big]\ .$$

- This is the optimal estimator in terms of MSE.

LLSE Estimator

- The linear least-squares error (LLSE) estimator is

$$\begin{aligned} \hat{x}_{\text{LLSE}}(y) &= \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\big(y - \mathbb{E}[Y]\big) \\ &= \mathbb{E}[X] + \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}\big(y - \mathbb{E}[Y]\big) \end{aligned}$$

- Attains the minimum MSE amongst all linear estimators.

$$\mathsf{MSE}_{\text{LLSE}} = \mathsf{Var}[X] - \frac{\big(\mathsf{Cov}[X,Y]\big)^2}{\mathsf{Var}[Y]} = \sigma_X^2(1 - \rho_{X,Y}^2)$$

- For jointly Gaussian X and Y, $\hat{x}_{\text{LLSE}}(y) = \hat{x}_{\text{MMSE}}(y)$.

Vector Estimation

- We observe a random vector $\underline{Y}$ and want to estimate an unobserved random vector $\underline{X}$ using an estimator $\underline{\hat{x}}(\underline{Y})$.
- Mean-Squared Error: $\mathsf{MSE} = \mathbb{E}\big[\big(\underline{X} - \underline{\hat{x}}(\underline{Y})\big)^{\mathsf{T}}\big(\underline{X} - \underline{\hat{x}}(\underline{Y})\big)\big]$
- The vector MMSE estimator is

$$\underline{\hat{x}}_{\text{MMSE}}(\underline{y}) = \mathbb{E}\big[\underline{X} | \underline{Y} = \underline{y}\big]$$

- The vector MMSE estimator attains the optimal MSE.
- The vector LLSE estimator is

$$\underline{\hat{x}}_{\text{LLSE}}(\underline{y}) = \mathbb{E}[\underline{X}] + \boldsymbol{\Sigma}_{\underline{X},\underline{Y}}\boldsymbol{\Sigma}_{\underline{Y}}^{-1}\big(\underline{y} - \mathbb{E}[\underline{Y}]\big)$$

  where $\boldsymbol{\Sigma}_{\underline{Y}}$ is the covariance matrix of $\underline{Y}$ and $\boldsymbol{\Sigma}_{\underline{X},\underline{Y}}$ is the cross-covariance matrix

$$\boldsymbol{\Sigma}_{\underline{X},\underline{Y}} = \mathbb{E}\big[(\underline{X} - \mathbb{E}[\underline{X}])(\underline{Y} - \mathbb{E}[\underline{Y}])^{\mathsf{T}}\big]\ .$$

- The vector LLSE estimator attains the optimal MSE amongst all linear estimators.
- If $\begin{bmatrix}\underline{X} \\ \underline{Y}\end{bmatrix}$ is a Gaussian vector, $\underline{\hat{x}}_{\text{LLSE}}(\underline{y}) = \underline{\hat{x}}_{\text{MMSE}}(\underline{y})$.

8. Sums of Random Variables

- Consider n random variables $X_1, X_2, \ldots, X_n$.
- We are often interested in the behavior of the sum $S_n = \sum_{i=1}^n X_i$ or the **sample mean** $M_n = \dfrac{1}{n}\sum_{i=1}^n X_i$.
- **Expected Value of the Sum:** $\mathbb{E}\big[S_n\big] = \sum_{i=1}^n \mathbb{E}[X_i]$
- **Variance of the Sum:** $\mathsf{Var}\big[S_n\big] = \sum_{i=1}^n\sum_{j=1}^n \mathsf{Cov}[X_i, X_j]$
- Random variables $X_1, \ldots, X_n$ are said to be **independent and identically distributed (i.i.d.)** if they are independent and all X_i have the same marginal distribution, which is a PMF $P_X(x)$ in the discrete case and a PDF $f_X(x)$ in the continuous case.
- For i.i.d. $X_1, \ldots, X_n$, we have that $\mathbb{E}\big[S_n\big] = n\mathbb{E}[X]$, $\mathsf{Var}\big[S_n\big] = n\mathsf{Var}[X]$, $\mathbb{E}\big[M_n\big] = \mathbb{E}[X]$, $\mathsf{Var}\big[M_n\big] = \mathsf{Var}[X]/n$.

Laws of Large Numbers

- **Weak Law of Large Numbers:** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean μ and sample mean M_n. For any $\epsilon > 0$, $\lim_{n\to\infty}\mathbb{P}\big[|M_n - \mu| > \epsilon\big] = 0$.
- **Strong Law of Large Numbers:** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean μ and sample mean M_n. Then, $\mathbb{P}\big[\lim_{n\to\infty} M_n = \mu\big] = 1$.

Central Limit Theorem

- **Central Limit Theorem:** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean μ and finite variance σ^2. The CDF of $Y_n = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}}$ converges to the standard normal CDF, $\lim\limits_{n\to\infty} F_{Y_n}(y) = \Phi(y)$.
- For i.i.d. random variables with finite mean and variance, then $F_{Y_n}(y) \approx \Phi(y)$ is a good approximation for $n \geq 30$.

9. Statistics

- Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ and variance $\mathsf{Var}[X_i] = \sigma^2$.
- The **sample mean** is $\hat{\mu} = M_n = \dfrac{1}{n}\sum_{i=1}^n X_i$.
- M_n is unbiased estimator for the mean, $\mathbb{E}[M_n] = \mu$, with variance $\mathsf{Var}[M_n] = \sigma^2/n$.
- The **sample variance** is $\hat{\sigma}^2 = V_n = \dfrac{1}{n-1}\sum_{i=1}^n \big(X_i - M_n\big)^2$.
- V_n is an unbiased estimator for the variance, $\mathbb{E}[V_n] = \sigma^2$.
- If $Z_1, \ldots, Z_n$ are i.i.d. Gaussian$(0,1)$ random variables, then $Y = \sum_{i=1}^n Z_i^2$ is a **chi-squared random variable with n degrees-of-freedom,** $Y \sim \chi^2_n$.
- If Z is a Gaussian$(0,1)$ random variable, Y is a chi-squared random variable with n degrees-of-freedom, and Y and Z are independent, then $W = Z\sqrt{n/Y}$ has a **Student's t-distribution with n degrees-of-freedom,** $W \sim T_n$. CDF: $F_{T_n}(t)$. PDF: Symmetric about 0.

Confidence Intervals for the Mean

- Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ, variance σ^2, sample mean M_n, and sample variance V_n.
- $\big[M_n \pm \epsilon\big]$ is a **confidence interval** for the mean with **confidence level** $1 - \alpha$ if $\mathbb{P}[\mu - \epsilon \leq M_n \leq \mu + \epsilon] = 1 - \alpha$.

Confidence Interval: Known Variance

- When to use: Variance is known *or* $n > 30$ samples.
- Set $\epsilon = \sigma Q^{-1}(\alpha/2)/\sqrt{n}$
- If the variance σ^2 is unknown *and* we have $n > 30$ samples, substitute σ^2 with the sample variance V_n.
- $Q^{-1}(0.05) = 1.64$, $Q^{-1}(0.025) = 1.96$, $Q^{-1}(0.005) = 2.57$

Confidence Interval: Unknown Variance

- When to use: Variance is unknown *and* $n \leq 30$ samples.
- Set $\epsilon = -\sqrt{V_n} F^{-1}_{T_{n-1}}(\alpha/2)/\sqrt{n}$ where $F_{T_{n-1}}(t)$ is the CDF for a Student's t-distribution with $n-1$ degrees-of-freedom.

Significance Testing

- Only have a probability model for the **null hypothesis** H_0.
- The **significance level** $0 \leq \alpha \leq 1$ is used to determine when to **reject the null hypothesis**.
- Given a statistic calculated from the dataset, the **p-value** is the probability of observing a value at least this extreme under the null hypothesis.
  - If p-value $< \alpha$, then reject the null hypothesis.
  - If p-value $\geq \alpha$, then fail to reject the null hypothesis.

One-Sample Z-Test

- Null Hypothesis: $X_1, \ldots, X_n$ is i.i.d. Gaussian(μ, σ^2).
- When to use: Variance σ^2 is known *or* $n > 30$ samples.
- Informally, is the mean not equal to μ?
  1. Calculate the sample mean M_n.
  2. Z-statistic: $Z = \sqrt{n}(M_n - \mu)/\sigma$.
  3. p-value $= 2\Phi(-|Z|)$.
- If the variance σ^2 is unknown *and* we have $n > 30$ samples, substitute σ^2 with the sample variance V_n.
- $2\Phi(-1.64) = 0.1$, $2\Phi(-1.96) = 0.05$, $2\Phi(-2.57) = 0.01$

One-Sample T-Test

- Null Hypothesis: $X_1, \ldots, X_n$ is i.i.d. Gaussian(μ, σ^2).
- When to use: Variance σ^2 is unknown *and* $n \leq 30$ samples.
- Informally, is the mean not equal to μ?
  1. Calculate the sample mean M_n and variance V_n.
  2. T-statistic: $T = \sqrt{n}(M_n - \mu)/\sqrt{V_n}$.
  3. p-value $= 2F_{T_{n-1}}(-|T|)$.

Two-Sample Z-Test

- Null Hypothesis: $X_1, \ldots, X_{n_1}$ is i.i.d. Gaussian(μ, σ_1^2) and $Y_1, \ldots, Y_{n_2}$ is i.i.d. Gaussian(μ, σ_2^2).
- When to use: Variances σ_1^2 and σ_2^2 are known *or* $\min(n_1, n_2) > 30$.
- Informally, do the datasets have the same mean?
  1. Calculate the sample means $M_{n_1}^{(1)}$ and $M_{n_2}^{(2)}$.
  2. Z-statistic: $Z = \left(M_{n_1}^{(1)} - M_{n_2}^{(2)}\right) \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
  3. p-value $= 2\Phi(-|Z|)$.
- If the variances σ_1^2, σ_2^2 are unknown *and* we have $\min(n_1, n_2) > 30$ samples, substitute σ_1^2 with the sample variance $V_{n_1}^{(1)}$ and σ_2^2 with the sample variance $V_{n_2}^{(2)}$.
- $2\Phi(-1.64) = 0.1$, $2\Phi(-1.96) = 0.05$, $2\Phi(-2.57) = 0.01$

Two-Sample T-Test

- Null Hypothesis: $X_1, \ldots, X_{n_1}$ is i.i.d. Gaussian(μ, σ^2) and $Y_1, \ldots, Y_{n_2}$ is i.i.d. Gaussian(μ, σ^2). The mean μ is unknown.
- When to use: (Equal) variance σ^2 is unknown *and* $\min(n_1, n_2) \leq 30$.
- Informally, do the datasets have the same mean?
  1. Calculate the sample means $M_{n_1}^{(1)}, M_{n_2}^{(2)}$, sample variances $V_{n_1}^{(1)}, V_{n_2}^{(2)}$, and the pooled sample variance $\hat{\sigma}^2 = \big((n_1-1)V_{n_1}^{(1)} + (n_2-1)V_{n_2}^{(2)}\big)/\big(n_1+n_2-2\big)$.
  2. T-statistic: $T = \left(M_{n_1}^{(1)} - M_{n_2}^{(2)}\right) \Big/ \sqrt{\hat{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$.
  3. p-value $= 2F_{T_{n_1+n_2-2}}(-|T|)$.

10. Machine Learning

- We focused on **binary classification** where the goal is to decide between two hypotheses, but we do not have access to the underlying probability model.
- Instead, we have a **dataset** with n samples. The i^{th} sample $(\underline{X}_i, Y_i)$ consists of a feature vector $\underline{X}_i$ and a label Y_i.
- We use this dataset to come up with a **classifier** $D(\underline{x})$, which is a function that maps any possible observation vector $\underline{x}$ into a guess of its label, $+1$ or -1.
- To make sure we are not overfitting, we split our dataset into non-overlapping **training** and **test** datasets.
- The training set is used to construct our classifier $D(\underline{x})$ and the test set can only be used to evaluate its performance.
- **Training Error** = fraction misclassified training examples.
- **Test Error** = fraction misclassified test examples.
- The **closest average** classifier computes the sample mean vectors $\hat{\underline{\mu}}_+$ and $\hat{\underline{\mu}}_-$ for each label. Given input $\underline{x}$, it computes the distances to $\hat{\underline{\mu}}_+$ and $\hat{\underline{\mu}}_-$ and chooses the label with the smallest distance.

$$D_{\text{avg}}(\underline{x}) = \begin{cases} +1, & \|\underline{x} - \hat{\underline{\mu}}_+\| \leq \|\underline{x} - \hat{\underline{\mu}}_-\|, \\ -1, & \text{otherwise} \end{cases}$$

- The **nearest neighbor** classifier outputs the label of the closest training example as its guess.

$$D_{\text{NN}}(\underline{x}) = Y_{\text{train},i_{\text{closest}}} \quad i_{\text{closest}} = \underset{i=1,\ldots,n_{\text{train}}}{\arg\min} \|\underline{x} - \underline{X}_{\text{train},i}\|$$

- The **LDA** classifier assumes the observation vectors are Gaussian vectors, with different mean vectors $\hat{\underline{\mu}}_+$ and $\hat{\underline{\mu}}_-$ and the same covariance matrix $\hat{\mathbf{\Sigma}}$.

$$D_{\text{LDA}}(\underline{x}) = \begin{cases} +1 & 2\,(\hat{\underline{\mu}}_+ - \hat{\underline{\mu}}_-)^{\mathsf{T}} \hat{\mathbf{\Sigma}}^{-1} \underline{x} \;\geq\; \hat{\underline{\mu}}_+^{\mathsf{T}} \hat{\mathbf{\Sigma}}^{-1} \hat{\underline{\mu}}_+ - \hat{\underline{\mu}}_-^{\mathsf{T}} \hat{\mathbf{\Sigma}}^{-1} \hat{\underline{\mu}}_-, \\ -1 & \text{otherwise.} \end{cases}$$

- The **QDA** classifier assumes the observation vectors are Gaussian vectors, with different mean vectors $\hat{\underline{\mu}}_+$ and $\hat{\underline{\mu}}_-$ and the covariance matrices $\hat{\mathbf{\Sigma}}_+$ and $\hat{\mathbf{\Sigma}}_-$.
- To avoid overfitting, we can use **PCA** to perform **dimensionality reduction:** $\underline{X}^{\mathsf{T}}_{\text{reduced}} = (\underline{X}^{\mathsf{T}} - \hat{\underline{\mu}}^{\mathsf{T}}_{\text{train}})\mathbf{V}_k$ where the columns of $\mathbf{V}_k$ are the k eigenvectors of $\hat{\mathbf{\Sigma}}_{\text{train}}$ corresponding to the largest k eigenvalues.

11. Markov Chains

- A **Markov chain** is a sequence of discrete random variables $X_0, X_1, X_2, \ldots$ such that, given the history $X_0, \ldots, X_n$, the next state X_{n+1} only depends on the current state X_n,

$$P_{X_{n+1}|X_n,\ldots,X_0}(x_{n+1}|x_n, \ldots, x_0) = P_{X_{n+1}|X_n}(x_{n+1}|x_n)$$

- We assume the range is finite $R_X = \{1, \ldots, K\}$.
- The **transition probabilities** P_{jk} are the probabilities of moving from state j to state k in one time step. We assume the Markov chain is **homogeneous**, $P_{X_{n+1}|X_n}(k|j) = P_{jk}$.
- The **n-step transition probabilities** $P_{jk}(n)$ are the probabilities of moving from state j to state k in exactly n time steps, $P_{jk}(n+m) = \sum_{i=1}^{K} P_{ji}(n) P_{ik}(m)$.
- The **state transition matrix** is $\mathbf{P} = \begin{bmatrix} P_{11} & \cdots & P_{1K} \\ \vdots & \ddots & \vdots \\ P_{K1} & \cdots & P_{KK} \end{bmatrix}$
- Row index is for current state, column index for next state.
- The **probability state vector** is $\underline{p}_t = \begin{bmatrix} P_{X_t}(1) \\ \vdots \\ P_{X_t}(K) \end{bmatrix}$.
- Moving forward one time step: $\underline{p}_{t+1} = \mathbf{P}^{\mathsf{T}} \underline{p}_t$.
- Moving forward n time steps: $\underline{p}_{t+n} = (\mathbf{P}^n)^{\mathsf{T}} \underline{p}_t$.

State Classification

- State k is **accessible** from state j if it is possible to reach state k starting from state j in zero or more time steps. Notation: $j \to k$ (State j is always accessible from itself.)
- States j and k **communicate** if $j \to k$ and $k \to j$. Notation: $j \leftrightarrow k$. (State j always communicates with itself.)
- A **communicating class** C is a subset of states such that if $j \in C$, then $k \in C$ if and only if $j \leftrightarrow k$.
- A Markov chain is **irreducible** if all of its states belong to a single communicating class.
- A communicating class C is **transient** if there are states $j \in C$ and $k \notin C$ such that $j \to k$ but $k \not\to j$.
- A communicating class that is not transient is **recurrent.**
- The **period** d of a state j is the greatest common divisor of the length of all cycles from j back to itself.
- All states in a communicating class have the same period.
- If the period is 1, then the state is called **aperiodic**. A Markov chain is aperiodic if all its states are aperiodic.

Limiting Probability State Vector

- For an irreducible, aperiodic Markov chain, there is a **unique limiting state probability vector** $\underline{\pi} = \lim_{t\to\infty} \underline{p}_t$ satisfying the following properties:
  - Normalization: $\sum_{j=1}^{K} \pi_j = 1$
  - Any initial state $\underline{p}_0$ will converge to $\underline{\pi}$.
  - **Steady-State Distribution:** $\underline{\pi} = \mathbf{P}^{\mathsf{T}} \underline{\pi}$.
- When $\underline{\pi}$ exists, we can solve for it using linear equations from $\underline{\pi} = \mathbf{P}^{\mathsf{T}} \underline{\pi}$ and $\sum_{j=1}^{K} \pi_j = 1$.
- If there is only one recurrent communicating class and it is aperiodic, then there is still a unique limiting state probability vector. Find by first setting π_j for all transient states to 0.