CSCI 6521: Advanced Machine Learning I
Study Guide for Test#1

(Please don't distribute this study guide.
The guide is for your study purpose only)

01. (a) What is the expected value? (b) Show that Var[x] = E[x^2] – (E[x])^2, where E is the expected value.

Ans:
(a) Expected value (E): Expected value is the value that we expect to see on average.

In probability theory, the expected value for a random value X with possible outcomes {x_1, x_2, ... , x_n} and with respective probabilities {p_1, p_2, …, p_n} can be calculated as:

$$E(X) = \sum_{i=1}^{n} x_i p_i = \mu$$

So, the above Equation is the
Expected Value,
Mean, or
Average
of the random variable X.

For example, for a fair die, if all six sides are equally likely (i.e., the probability of the outcome of any one side is $1/6$), then the expected value can be computed as:

$$E = (1\times\frac{1}{6}) + (2\times\frac{1}{6}) + (3\times\frac{1}{6}) + (4\times\frac{1}{6}) + (5\times\frac{1}{6}) + (6\times\frac{1}{6}) = \frac{21}{6} = 3.5$$

Therefore, if we play with the fair die for a long time, collect the outcomes and then take the average of the outcomes, theoretically, the value will be 3.5.

(b) $Var[x] = \sum_{x\in X}(x-\mu)^2 P(x)$

$$= (x_1-\mu)^2 P(x_1) + (x_2-\mu)^2 P(x_2) + ... + (x_n-\mu)^2 P(x_n)$$

$$= [x_1^2 P(x_1) + x_2^2 P(x_2) + ... + x_n^2 P(x_n)]$$
$$- 2\mu[x_1 P(x_1) + x_2 P(x_2) + ... + x_n P(x_n)]$$
$$+ \mu^2[P(x_1) + P(x_2) + ... + P(x_n)]$$

$$= E[x^2] - 2\mu\, E[x] + \mu^2 \qquad [\because P(x_1) + P(x_2) + ... + P(x_n) = 1 \quad]$$

$$= E[x^2] - 2\mu\mu + \mu^2 \qquad [\because E[x] = \mu]$$
$$= E[x^2] - 2\mu^2 + \mu^2$$
$$= E[x^2] - \mu^2$$
$$= E[x^2] - (E[x])^2 \qquad [\because E[x] = \mu]$$

02. Draw a figure to show the relationship among *Prediction Error, Model Complexity, Bias, Variance, Training dataset,* and *Test dataset.*

Ans:

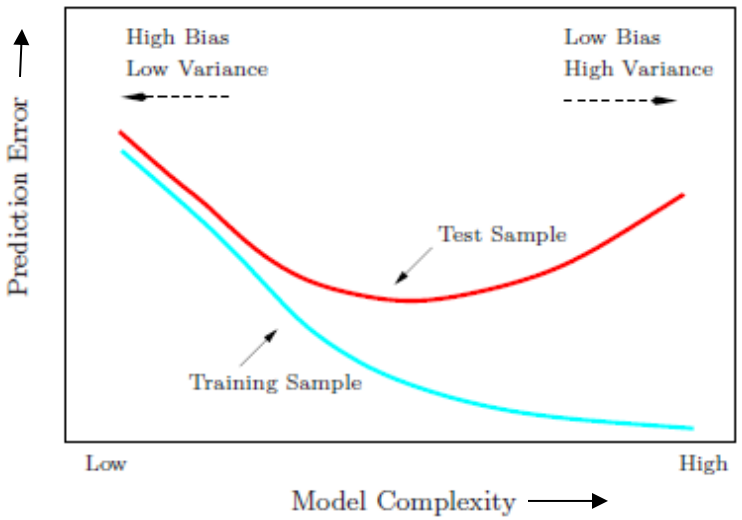

Figure: Test and training error as a function of model complexity.

03. Derive the bias-variance decomposition relationship for Mean Squared Error (MSE).

Ans:

Mean Squared Error (MSE): The MSE of a predictor is the mean of the square of the difference between the predicted value and the true value of the output being predicted. The MSE is also regarded as a *Risk Function*.

The MSE of a predictor *f*, when it predicts $\hat{Y} = f(X)$ for *Y*, can be written as,

$$MSE(\hat{Y}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

We can also write in terms of expected value as:

$$MSE(\hat{Y}) = E[(\hat{Y} - Y)^2]$$

Bias: The bias of a predictor is the difference between the predictor's expected value and the true value of the parameter.

Therefore,

$$Bias = (E[\hat{Y}] - Y)$$

Or, $$Bias = (\mu - Y) \qquad [\because mean, \mu = E[\hat{Y}]]$$

Now the variance (σ^2) is the square of the standard deviation σ.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \mu)^2} = \sqrt{E[(\hat{Y} - \mu)^2]}$$

$$\sigma^2 = E[(\hat{Y} - \mu)^2]$$

Now, we show, $MSE(\hat{Y}) = E[(\hat{Y} - Y)^2]$

$$= E[((\hat{Y} - \mu) + (\mu - Y))^2]$$

$$= E[(\hat{Y} - \mu)^2 + 2(\hat{Y} - \mu)(\mu - Y) + (\mu - Y)^2]$$

$$= E[(\hat{Y} - \mu)^2] + E[(\mu - Y)^2] \quad [\because E[\mu] = \mu, \mu = E[\hat{Y}], \therefore E(\hat{Y} - \mu) = 0]$$

$$= \sigma^2 + E[(\mu - Y)^2] \qquad [\because Var, \sigma^2 = E[(\hat{Y} - \mu)^2]]$$

$$= \sigma^2 + (\mu - Y)^2 \quad [\because \text{E of a constant will remain the same}]$$

$$= \sigma^2 + (Bias)^2$$

Therefore,

$$MSE = Var + bias^2$$

The variance measures the extent to which the solutions for individual data sets vary around their average, and hence this measures the extent to which the function *f*(*X*) is sensitive to the particular choice of data set.

The bias represents the extent to which the average prediction over all data sets differs from the desired regression function. The MSE can be determined from *Var* and *bias* alternatively.

04. (a) What is the law of the total probability? (b) Using the "law of the total probability", solve the following problem: "Both Tamiflu and Relenza are drugs that work against influenza. Let us assume, our study at UNO found that Tamiflu remains effective for 12 months in 80% of cases, and Relenza remains effective for 12 months in 90% of cases. If UNO purchases 30%

Tamiflu and 70% Relenza to vaccinate its staff, how likely is it that staff with remain influenza free for 12 months?" (**see lecture note**)

Ans:

(a) The *Law of the Total Probability* states that if an event A can occur in *n* different ways: $A_1, A_2, ..., A_n$ and if these *n* subevents are mutually exclusive, i.e., they cannot occur at the same time, then the probability of A occurring is the sum of the probabilities of the subevents A_i. Specifically, the random variable y can assume the value of y in *n* different ways with *x*: $x_1, x_2, \ldots, x_n$.

Because these probabilities are mutually exclusive, it follows from the *Law of the Total Probability* that $P(y)$ is the sum of the joint probability $P(x, y)$ over all possible values of *x*. That is,

$$P(y) = P(y \mid x_1)P(x_1) + P(y \mid x_2)P(x_2) + ... + P(y \mid x_n)P(x_n)$$

(b) Using the law of total probability, we can write:

$P(y) = P(y|x_1)\, P(x_1) + P(y|x_2)\, P(x_2)$
$= (0.80)\ (0.30) + (0.90)\ (0.70)$
$= 0.87$

Therefore, there is an 87% chance that staff will remain influenza-free for 12 months.

05. (a) Write down the Bayes Equation and name different parts of it. (b) Using the "Bayes Equation", solve the following problem: Suppose we have invented an influenza tester. Now, say 5% of the population is sneezing due to the cold season. Running test on sneezing people, 90% test returned positive for influenza testing. Given a person is NOT sneezing, the test comes out positive 15% of the time. For a positive detection, what is the chance that the person is sneezing?
Ans:

(a) Bayes Equation:

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{\sum_{x \in X} P(y \mid x)P(x)}$$

Name of different parts:

$$posterior = \frac{likelihood \times prior}{evidence}$$

(b) Here,

- 5% of the population is **S**neezing, $p(S) = 0.05$ [*prior*]
- The test returns **T**rue given a person **S**neezing for 90% cases, $p(T \mid S) = 0.9$ [*likelihood*]
- The test is True for non-Sneezing people in 15% cases, $p(T \mid \bar{S}) = 0.15$

We are calculating the chance that given True or positive detection of influenza, what is the change that the person is sneezing, i.e., $p(S \mid T) = ?$ [*posterior?*]

Following Bayes rule, we can write,

$$p(S \mid T) = \frac{p(T \mid S)p(S)}{p(T)}$$

$$= \frac{p(T \mid S)p(S)}{\sum_S p(T \mid S)p(S)}$$

$$= \frac{p(T \mid S)p(S)}{p(T \mid S)p(S) + p(T \mid \bar{S})p(\bar{S})}$$

$$= \frac{p(T \mid S)p(S)}{p(T \mid S)p(S) + p(T \mid \bar{S})[1 - p(S)]}$$

$$= \frac{0.90 \times 0.05}{0.90 \times 0.05 + 0.15 \times [1 - 0.05]}$$

$$= \frac{0.90 \times 0.05}{0.90 \times 0.05 + 0.15 \times 0.95}$$

$$= \frac{0.045}{0.1875} = 0.24 \text{ or, } 24\% \text{ chance.}$$

Finding a person positive for influenza, the change that the person is sneezing is 24%,

Or,

We can also say, the degree of belief that the person is sneezing is 24%.

06. What are the properties of a (univariate) Gaussian / Normal distribution *N* $(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$**?**

Ans: Here, variable $x \in (-\infty, \infty)$

Mean $\mu \in (-\infty, \infty)$ and

Variance $\sigma^2 > 0$

Mean = Median = Mode

And, $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = 1$

The equation is bell-shaped and symmetric.

07. Using maximum likelihood estimation, for a given data D{ $x_1, x_2, ...x_N$ } where iid is followed and pdf is Gaussian [*P*(*x*) = $\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$], estimate the best mean (μ).

Ans:

Let us assume, the given data D = { $x_1, x_2, ...x_N$ }, iid is followed, pdf is Gaussian. We want to estimate the best mean (μ) using MLE.

Since pdf is Gaussian, we write, $P(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

$$\text{Since, iid, } L(\mu) = \prod_{i=1}^{N} P(x_i) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \quad \ldots \qquad \text{(A)}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x_1-\mu)^2}{2\sigma^2}\right\}\right) \times \left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x_2-\mu)^2}{2\sigma^2}\right\}\right) \times \ldots \left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x_N-\mu)^2}{2\sigma^2}\right\}\right)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \times \exp\left\{-\frac{(x_1-\mu)^2}{2\sigma^2} - \frac{(x_2-\mu)^2}{2\sigma^2} - \ldots - \frac{(x_N-\mu)^2}{2\sigma^2}\right\}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \times e^{-\left\{\frac{(x_1-\mu)^2}{2\sigma^2} + \frac{(x_2-\mu)^2}{2\sigma^2} + \ldots + \frac{(x_N-\mu)^2}{2\sigma^2}\right\}}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \times e^{-\frac{1}{2\sigma^2}\left\{(x_1-\mu)^2 + (x_2-\mu) + \ldots + (x_N-\mu)^2\right\}}$$

To maximize the Equation (A), we can only do that by minimizing $\left\{(x_1-\mu)^2 + (x_2-\mu) + \ldots + (x_N-\mu)^2\right\}$, right?

Now, expanding the target, we can write our goal for MLE as,

$$f(\mu)_{\min} = N\mu^2 - 2\mu(x_1 + x_2 + \ldots + x_N) + (x_1^2 + x_2^2 + \ldots + x_N^2) \qquad \text{(B)}$$

Equation (B) is a quadratic function, and the coefficient of μ^2, i.e., *N* is positive. Thus the minimum exists.

Therefore to get minimum, we can differentiate $f(\mu)$ and then set 0, i.e., $f'(\mu) = 0$

$$\Rightarrow\ 2.N.\mu - 2(x_1 + x_2 + ... + x_N) = 0$$

$$\Rightarrow\ \mu = \frac{(x_1 + x_2 + ... + x_N)}{N}$$

Therefore, our estimated best mean using MLE is actually the arithematic average of the given data.

Note: Alternatively, you can also use log-likelihood to answer the above question.

08. Describe Naïve Bayes Classifier?

Ans**:**

The Naïve Bayes classifier is based on the Bayes rule. Naïve Bayes is popular being simple, intuitive, and works well in the higher dimensional space. More generally, assume, for a given set of variables, $X = \{x_1, x_2, ..., x_d\}$, and class sets $C_i = \{C_1, C_{2,}, ..., C_k\}$, from Bayes rule, we can write Equation (1) ignoring the evidence as:

$$P(C_i \mid x_1, x_2, ..., x_d) \equiv p(x_1, x_2, ..., x_d \mid C_i)P(C_i) \qquad (1)$$

Here, Naïve Bayes assumes that the conditional probabilities of the variables are statistically independent (this is a "naïve" assumption, thus known as "Idiot's Bayes, however, practically naïve Bayes classifiers often outperform far more sophisticated alternatives). This likelihood is computed as products:

$$P(X \mid C_i) \equiv \prod_{j=1}^{d} p(x_j \mid C_i) \qquad (2)$$

Without the naïve assumption or, in other words, without the resulting Equation (2), the computation would have been expensive. This could have been expanded using the chain rule as:

$$\begin{aligned} P(C_i \mid x_1, x_2, ..., x_d) &\equiv p(x_1, x_2, ..., x_d \mid C_i)P(C_i) \\ &= p(x_2, ..., x_d \mid C_i, x_1)P(x_1 \mid C_i)P(C_i) \\ &= p(x_3, ..., x_d \mid C_i, x_1, x_2)P(x_2 \mid C_i, x_1)P(x_1 \mid C_i)P(C_i) \\ &= \ldots \end{aligned}$$

Finally, we can summarize from Equation (1) and (2):

$$P(C_i \mid X) \equiv P(C_i) \prod_{j=1}^{d} p(x_j \mid C_i) \qquad (3)$$

We can use Equation (3) for the classification problem and can pick the class as:

$$C_i = \arg\max_{c_i} P(c_i) \prod_{j=1}^{d} p(x_j \mid c_i) \qquad (4)$$

Finally, we pick the class as an answer to our query, for which the value is found maximum.

09. What is Laplace Smoothing or Additive Smoothing used for?
Ans: (see lecture notes)

10. Explain the central limit theorem? (see lecture notes + complete it by yourself)

11. Explain Bernoulli distribution?
Ans: (see lecture notes)

12. Prove that for the variables *x* and *y*, the covariance, σ_{xy}, can be computed as: $\sigma_{xy} = E[xy] - \mu_x \mu_y$.

Ans: We know, $\sigma_{xy} = E[(x - \mu_x)(y - \mu_y)]$

$= E[(xy - x\mu_y - y\mu_x + \mu_x \mu_y)]$

$= E[xy] - E[x]\mu_y - E[y]\mu_x + \mu_x \mu_y$

$= E[xy] - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y$

$= E[xy] - 2\mu_x \mu_y + \mu_x \mu_y$

$= E[xy] - \mu_x \mu_y$

Therefore, $\sigma_{xy} = E[xy] - \mu_x \mu_y$

Or, $\sigma_{xy} = E[xy] - E[x]E[y]$

13. (a) Define and describe the generative model. (b) Describe the framework of the generative model.

Ans:
(a) A generative model describes how a dataset is generated in terms of a probabilistic model. By sampling from this model, we are able to generate new data.

For example, suppose we have a dataset containing images of horses. We may wish to build a model that can generate a new image of a horse that has never existed but still looks real because the model has learned the general rules that govern the appearance of a horse. This is the kind of problem that can be solved using generative modeling.

A generative model must also be probabilistic rather than deterministic. If our model is merely a fixed calculation, such as taking the average value of each pixel in the dataset, it is not generative because the model produces the same output every time. The model must include a stochastic (random) element that influences the individual samples generated by the model.

(b) The framework of the generative model is described as follows:

We have a dataset of observations **X**.

- We assume that the observations have been generated according to some unknown distribution, p_{data}.
- A generative model p_{model} tries to mimic p_{data}. If we achieve this goal, we can sample from p_{model} to generate observations that appear to have been drawn from p_{data}.
- We are impressed by p_{model} if:
  - Rule 1: It can generate examples that appear to have been drawn from p_{data}.
  - Rule 2: It can generate examples that are suitably different from the observations in **X**. In other words, the model shouldn't simply reproduce things it has already seen.

14. Given the univariate Gaussian distribution, $N(x \mid \mu, \sigma^2) = p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ **, build the multivariate Gaussian distribution:** $p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$**. To make the derivation simple, you may assume the covariance matrix as a diagonal matrix (see the lecture note).**

--- x ---