The use of the internet, calculators, Prof B.’s homework solutions, or a cheat sheet is not allowed. The test is open book and open note. If using the computer for a book, you must clearly indicate that your wireless is off. If you are caught accessing the internet, you will get a zero on the test.

Attempt all problems and clearly indicate answers. Solutions given with little or no justification may receive little or no credit.

“I pledge on my honor that I have neither given nor received unauthorized aid on this assignment.”

Signature: ________________________________
1. (10 points) True/False questions. Defend your answer.

(i) We can get multiple local optima solutions if we solve a linear regression problem by minimizing the sum of squared errors.

<table>
<thead>
<tr>
<th>True</th>
<th>False</th>
</tr>
</thead>
</table>

**Solution:** False. The OLS solution is unique and is the MLE solution. Since the problem is convex, we know that this is the global maximum and that there are no other maxima.

(ii) When the hypothesis space is richer, over fitting is more likely.

<table>
<thead>
<tr>
<th>True</th>
<th>False</th>
</tr>
</thead>
</table>

**Solution:** True. The richer the hypothesis space, the more likely the model is to capture variance in the data.

(iii) Zero correlation between two random variables implies that the two variables are independent.

<table>
<thead>
<tr>
<th>True</th>
<th>False</th>
</tr>
</thead>
</table>

**Solution:** False. Correlation captures only linear independence. The variables could be related such that \( y = x^2 \) and if \( x \) is centered around zero, the correlation of these would be (near) zero.

(iv) The perceptron algorithm does not converge if the training samples are not linearly separable.

<table>
<thead>
<tr>
<th>True</th>
<th>False</th>
</tr>
</thead>
</table>

**Solution:** True. The perceptron algorithm’s update rule is to adjust the parameters if a point is misclassified, so if it is not possible to correctly classify all training data (e.g. linearly separable) then the algorithm will not reach a stable point.

(v) The maximum likelihood model parameters \((\theta_1, \theta_2)\) can be learned using linear regression for the model \( y_i = \log(x_1^\theta_1 e^{\theta_2}) + \epsilon \) where \( \epsilon_i \sim N(0,\sigma^2) \) is iid noise and \( x_1 > 0 \).

<table>
<thead>
<tr>
<th>True</th>
<th>False</th>
</tr>
</thead>
</table>

**Solution:** True. \( y_i = \log(x_1^\theta_1 e^{\theta_2}) + \epsilon = \theta_1 \log(x_1) + \theta_2 \) is linear in \( \theta_1 \) and \( \theta_2 \) so it can be learned using linear regression.
2. (16 points) Short answers

(i) We defined penalized sum of squares, a loss function used in linear regularization, as

\[
PSS(\theta, \lambda) = (Y - \theta X)^2 + \lambda \left( \sum_{i=1}^{p} (\theta_i)^2 \right).
\]

What is the role of \( \lambda \)? What happens when \( \lambda \) is very large or very small?

**Solution:** \( \lambda \) is a shrinkage parameter and penalizes the model based on the size (or magnitude) of \( \beta \)'s. If it is very large the parameter estimations are pushed towards zero, if it is small, the solution becomes closer to the OLS solution.

(ii) Define likelihood, posterior, and prior. You may assume the model is a classification task.

**Solution:** Likelihood is the probability of the data given the label. Prior refers to our belief about how likely each label is before we've seen any data. The posterior includes influences of both the prior and the likelihood.

\[
P(y = c | x) = \frac{P(x | y = c)P(y = c)}{P(x)}
\]

The posterior here is \( P(y = c | x) \), the likelihood is \( P(x | y = c) \) and the prior is \( P(y = c) \), \( P(x) \) is a normalizing constant that relates to how likely the observed data \( x \) is.

(iii) When should you use cross-validation? What does it allow you to estimate more accurately?

**Solution:** Always. It allows for a better estimate of generalization error, or error on unseen data.

(iv) What is the naive assumption of Naive Bayes?

**Solution:** That the predictors/features of \( X \) are conditionally independent given the class label.
3. (8 points) Derive the MLE for 10 random samples of a Bernoulli trial with the general loss function defined as

\[ \mathcal{L}(D, \theta) = \prod_{i=1}^{N} \theta^{x_i}(1 - \theta)^{(1-x_i)}. \]

**Solution:** Take the log such that

\[ \ell(D, \theta) = \sum_{i=1}^{10} x_i \log(p) - \sum_{i=1}^{10} (1 - x_i) \log(1 - p) \]

\[ = 10\bar{x} \log(p) - 10(1 - \bar{x}) \log(1 - p) \]

Take the derivative with respect to \( p \)

\[ \frac{d\ell}{dp} = 10\bar{x} \frac{1}{p} - 10(1 - \bar{x}) \frac{1}{1 - p} \]

\[ 0 = \frac{10\bar{x}(1 - p) - 10p(1 - \bar{x})}{p(1 - p)} \]

\[ 10\bar{x}(1 - p) = 10(1 - \bar{x})p \]

\[ \bar{x} - \bar{x}p = p - \bar{x}p \]

\[ \bar{x} = p \]

So the MLE of \( p \) is \( \bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i \)
4. Bias variance trade-off

(i) (6 points) Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

<table>
<thead>
<tr>
<th>Model</th>
<th>Bias</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>low/hi</td>
<td>low/hi</td>
</tr>
<tr>
<td>Polynomial regression with degree 3</td>
<td>low/hi</td>
<td>low/hi</td>
</tr>
<tr>
<td>Polynomial regression with degree 10</td>
<td>low/hi</td>
<td>low/hi</td>
</tr>
</tbody>
</table>

Solution:

<table>
<thead>
<tr>
<th>Model</th>
<th>Bias</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>low/hi</td>
<td>low/hi</td>
</tr>
<tr>
<td>Polynomial regression with degree 3</td>
<td>low/hi</td>
<td>low/hi</td>
</tr>
<tr>
<td>Polynomial regression with degree 10</td>
<td>low/hi</td>
<td>low/hi</td>
</tr>
</tbody>
</table>

(ii) (8 points) For each part below indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(i) The sample size \( n \) is extremely large, and the number of predictors \( p \) is small.

Solution: A flexible method is better because we are less at risk of overfitting if we have lots of data and only a few relevant predictors. Note you could make an argument that an inflexible method is better since we don’t have a lot of predictors. This is not incorrect but, if we have lots of data, use it! It doesn’t get as much use in inflexible models.

(ii) The number of predictors \( p \) is extremely large, and the number of observations \( n \) is small.

Solution: Use an inflexible method. Since \( p \) is very large, it is easy to overfit or to incorporate predictors into the model that are not actually helpful in predicting the output (e.g. your model may capture noise/variance). Also for small \( n \) you are much more likely to see spurious relationships that aren’t actually present in the population.

(iii) The relationship between the predictors and response is highly non-linear.

Solution: Flexible models. Flexible models will allow you to capture different (non-linear) relationships. Unless you know exactly what the relationship between \( X \) and \( Y \) is and you choose a very inflexible model that happens to capture a specific non-linear relationship that is true in the world, a flexible hypothesis space is more likely to work better.

(iv) The variance of the error terms, i.e. \( \sigma^2 = \text{Var}(\epsilon) \) is extremely high.

Solution: Inflexible. You are very likely to find relationships that are just due to noise. Flexible models will also try to find and fit patterns in the irreducible noise which will cause high variance in the final model.
5. Multivariate Normals, LDA/QDA The class of Gaussian Discriminant Analysis models are generative models. The generative assumptions underlying the models are that

\[ y \sim \text{Bernoulli}(\phi), \]
\[ x|y = c \sim \text{MVN}(\mu_c, \Sigma_c). \]

(i) (4 points) How many parameters are in the QDA model assuming that we have \( p \) features for each \( x_i \) observation and \( n \) data points.

**Solution:** We know the model is generative so our model is fully specified by

\[ p(y|x) \propto p(x|y)p(y) \]

We know that \( y \) can take two values because it is distributed as a binomial. So we need only 1 parameter (the probability of success) to fully specify our prior. To specify the likelihood (for a single class), we need to estimate the mean for each class \((\mu_i|y = c)\) and the joint co-variance \( \Sigma_c \). We note that the covariance is symmetric but that the diagonal elements are unique. So the likelihood is specified by \( 2p + \frac{p^2 + p}{2} \times 2 \) since there are 2 classes. In total that means we need \( 1 + 2p + p^2 + p = p^2 + 3p + 1 \) parameters.

(ii) (6 points) Rewrite the model definition assuming Linear Discriminant Analysis. How many parameters in this model?

**Solution:**

\[ y \sim \text{Bernoulli}(\phi), \]
\[ x|y = 0 \sim \text{MVN}(\mu_0, \Sigma), \]
\[ x|y = 1 \sim \text{MVN}(\mu_1, \Sigma). \]

We can make a similar argument as above to show that the model has \( 1 + 2p + p^2 + p = p^2 + 3p + 1 \) parameters.

(iii) (2 points) What does MVN stand for? What is \( \mu \) and \( \Sigma \)? What does the \( \Sigma_{ij} \) represent?

**Solution:** MVN stands for multivariate normal. \( \mu \) is the vector of means for each predictor in \( x \), such that \( \mu_i \) is the population mean of predictor \( i \). \( \Sigma \) is a \( p \times p \) matrix such that each \( \Sigma_{ij} = \text{cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)] \) for each pair of predictors. Note that if \( i = j \), or along the diagonal of \( \Sigma \), we get the variance of each predictor \( i \).
6. Consider fitting a linear regression model for these data

\[ \begin{array}{c|ccc}
   x & -1 & 0 & 2 \\
   \hline
   y & 1 & -1 & 1 \\
\end{array} \]

(i) (4 points) Fit \( y_i = \beta_0 + \epsilon_i \) (intercept only model), find \( \beta_0 \) using MSE as our loss function.

**Solution:**

\[ L(\beta_0) = \frac{1}{N} \sum_{i=1}^{3} (y_i - \beta_0)^2 \]

\[ \frac{dL}{d\beta_0} = 2 \sum_{i=1}^{3} (y_i - \beta_0) = 0 \]

\[ 0 = (1 - \beta_0) + (-1 - \beta_0) + (1 - \beta_0) \\
   -1 = -3\beta_0 \\
   \frac{1}{3} = \beta_0 \]

(ii) (6 points) Fit \( y_i = \beta_1 x_i + \epsilon_i \) (linear regression without the intercept term), find \( \beta_1 \) using MSE as our loss function.

**Solution:**

\[ L(\beta_1) = \frac{1}{N} \sum_{i=1}^{3} (y_i - \beta_1 x_i)^2 \]

\[ \frac{dL}{d\beta_1} = 2 \sum_{i=1}^{3} (y_i - x_i \beta_1) x_i = 0 \]

\[ 0 = (1 - (-1)\beta_1)(-1) + (-1 + 0)(0) + (1 - (2)\beta_1)(2) \\
   = -1 - \beta_1 + 2 - 4\beta_1 \\
   \beta_1 = \frac{1}{5} \]
7. Consider the algorithm for gradient descent. The update rule is

\[ \theta_j := \theta_j - \alpha \nabla_{\theta_j} f(h_{\theta}(x)) \]

(i) (8 points) Explain how this update works. Discuss what each component of the update rule is and if/how this reaches the global minimum (you may assume that the problem is convex).

**Solution:** Starting at an arbitrary initialization for \( \theta \), this procedure allows us to find the local minima of a function by iteratively updating our estimate. We decide the direction to move by finding the gradient. The gradient will be zero at the minimum and thus the algorithm will stop when it reaches a minima. In the case of a convex problem, this local minima will be the global minima as well. \( \alpha \) controls the “step size” or how much we move in the direction of the gradient at any point in time. The term following the \( \alpha \) is the gradient with respect to the loss function we are using to estimate the parameters \( \theta \).

(ii) (2 points) Name two algorithms where we saw (a variant) of this gradient descent update rule.

**Solution:** LMS (least mean squared error), Logistic regression gradient descent, perceptron learning algorithm and back propagation are all algorithms that use a variation of gradient descent to iteratively update parameters.
8. (i) How difficult was this exam?

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(ii) Do you have any comments on the quality of the questions, the material tested, or on the number of questions?

(iii) How difficult is this course?

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(iv) Do you have any comments on the quality of the lectures, the material covered, or on the course more generally?