ECE 661 Spring 2025 Gradescope Quizzes

Introduction to Machine Learning for Engineers
Prof. Gauri Joshi and Prof. Carlee Joe-Wong

Problem 1: When we set $\alpha = 1$ and $\beta = 1$ in the Beta(α, β) distribution it reduces to this distribution:

- ◯ Uniform
- ◯ Bernoulli
- ◯ Gaussian
- ◯ Exponential

Solution: A. Since Beta$(x, \alpha, \beta) = C_{\alpha,\beta} x^{\alpha-1}(1-x)^{\beta-1}$, if $\alpha = \beta = 1$, Beta(x, α, β) becomes constant for all x. It must then be a uniform distribution.

Problem 2: For which of these prior distributions does the Maximum a Posteriori (MAP) reduce to the Maximum Likelihood Estimation (MLE)?

- ◯ Uniform
- ◯ Bernoulli
- ◯ Gaussian
- ◯ Exponential

Solution: A. A uniform prior gives no information about the parameter value, so MAP will give the same estimate as MLE.

Problem 3: Suppose you have data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ and the solution to linear regression on this data is $y = w_1 x + b_1$. Now suppose we have the dataset $(x^{(1)} + \alpha, y^{(1)} + \beta), \ldots, (x^{(n)} + \alpha, y^{(n)} + \beta)$ where $\alpha > 0, \beta > 0$ and $w_1\alpha \neq \beta$. The solution to the linear regression on this new dataset is $y = w_2 x + b_2$.

Please select the correct statement about w_1, w_2, b_1, b_2 below. Note that the statement should hold no matter what values α, β take on within the specified constraints.

- ◯ $w_2 = \alpha w_1/\beta; \quad b_2 = b_1/\beta$
- ◯ $w_2 = \beta w_1/\alpha; \quad b_2 = b_1 + \beta$
- ◯ $w_2 = w_1; \quad b_2 = b_1 - w_1\alpha$
- ◯ $w_2 = w_1; \quad b_2 = b_1 - w_1\alpha + \beta$

Solution: D. Using these values for w_2 and b_2, we predict the label $w_1\left(x^{(i)} + \alpha\right) + b_1 - w_1\alpha + \beta = w_1 x^{(i)} + b_1 + \beta$ for each data point i. Thus, the prediction error for data point i will be $w_1 x^{(i)} + b_1 + \beta - (y^{(i)} + \beta) = w_1 x^{(i)} + b_1 - y^{(i)}$, which is the same prediction error as for the original dataset.

Problem 4: Suppose we are given a dataset (x_n, y_n) for $n = 1, \ldots, N$, where x_n and y_n are scalars. We use polynomial regression to fit a function $f(x) = w_0 + w_1 x + \cdots + w_M x^M$ to the data. As we increase the degree of the polynomial M we get a model with

◯ Higher bias, lower variance

◯ Lower bias, higher variance

Solution: B. Increasing the degree of the polynomial gives a more complex model, which can more closely fit the training data (lowering the bias) but may increase the variance.

Problem 5: Which of the following statements are true about addition of the regularizer term $\lambda\|\mathbf{w}\|^2$ to the residual sum of squares objective function? More than one option can be true.

☐ The solution $\mathbf{w}$ with $\lambda > 0$ will have a higher training loss than the solution with $\lambda = 0$.

☐ The solution $\mathbf{w}$ with $\lambda > 0$ will have a lower training loss than the solution with $\lambda = 0$.

☐ The solution $\mathbf{w}$ with $\lambda > 0$ will have a larger L_2 norm than the solution with $\lambda = 0$.

☐ The solution $\mathbf{w}$ with $\lambda > 0$ will have a smaller L_2 norm than the solution with $\lambda = 0$.

Solution:

☐ True, since the solution with $\lambda = 0$ will by definition minimize the training loss.

☐ False, since the option above is true.

☐ False, since including the regularization term will reduce the L_2 norm of the optimal $\mathbf{w}$.

☐ True, see item above.

Problem 6: Suppose you use a simple bag-of-words Naive Bayes model to determine whether a document was written by an AI generator or a human. Given a set of documents as training data, you find that the prior probability of the "AI generator" class, $\pi_{\text{AI generator}}$, is 0.7, and that the conditional probability of the word "machine" given this class, $P(machine|\text{class} = \text{"AI generator"})$, is 0.02.

Now suppose you add five new documents to your training data, each of which was written by an AI generator. You re-compute the Naive Bayes model parameters on your training data with these additional documents. Which of the following are feasible updated values of $\pi_{\text{AI generator}}$ and $P(machine|\text{class} = \text{"AI generator"})$?

◯ $\pi_{\text{AI generator}} = 0.5, P(machine|\text{class} = \text{"AI generator"}) = 0.2$

◯ $\pi_{\text{AI generator}} = 0.75, P(machine|\text{class} = \text{"AI generator"}) = 0.3$

◯ $\pi_{\text{AI generator}} = 0.7, P(machine|\text{class} = \text{"AI generator"}) = 0.3$

◯ $\pi_{\text{AI generator}} = 0.8, P(machine|\text{class} = \text{"AI generator"}) = 0$

Solution: B. A and C are incorrect since more than 70% of our documents are AI-generated after the addition of new documents. D is incorrect since we have a nonzero probability of observing the word machine in an AI-generated document.

Problem 7: Consider a binary classification problem. Suppose that you train a naive Bayes model to solve this problem, using Laplacian smoothing with parameter α to compensate for features that are missing in one class. Recall that to do so, we pretend to have seen each feature α additional times in each class.

As you increase the Laplacian smoothing parameter α, which of the following will happen to your model?

◯ The model's predictions on a test dataset will become more similar to those made without Laplacian smoothing.

◯ The model will become equivalent to MAP estimation with a uniform prior distribution.

◯ The prior probabilities for each class will eventually approach 0.5.

◯ None of the above.

Solution: Depends on interpretation. B is correct if we interpret its statement as using a uniform prior only on the conditional probabilities. If we interpret this statement as using a uniform prior for all probabilities (both the conditional and class prior probabilities), then D is correct.

Problem 8: Suppose that you tried to train a logistic regression model on a dataset $\mathcal{D}$, but you found that the optimal model classified some of your training data points incorrectly. You tried to fix this by using nonlinear basis functions to transform the original input features x_1, x_2 into the new features x_1, x_2, x_1^2, x_2^2, but your model trained on the new features still classified some of the training data points incorrectly.

Which of the following changes might ensure that your learned model classifies all of the training data points correctly?

◯ Train the logistic regression model using the features $x_1, x_2, x_1 + x_2^2$.

◯ Introduce a regularization term into the objective function.

◯ Collect new training data points and add them to $\mathcal{D}$.

◯ None of the above.

Solution: D. A is incorrect since these features are linear combinations of the nonlinear features we already tried. B is incorrect since introducing a regularization term always reduces the training loss. C is incorrect since introducing new training data will increase the difficulty of correctly classifying all training data points.

Problem 9: Consider a classification dataset $\mathcal{D}$ with five possible class labels. Suppose you wish to use logistic regression to train a model that predicts the class to which a given data point $\mathbf{x}$ belongs. You decide to try both the one-vs-one and one-vs-all methods.

How many binary logistic regression models did you train for each method?

◯ 5 for one-vs-all, 10 for one-vs-one

◯ 10 for one-vs-all, 5 for one-vs-one

◯ 5 for one-vs-all, 5 for one-vs-one

◯ 5 for one-vs-all, 20 for one-vs-one

Solution: A. We train one model for each class for one-vs all (i.e., 5 models) and one model per pair of classes for one-vs-one. Since we have 5 classes, we have $\frac{5(5-1)}{2} = 10$ pairs of classes.

Problem 10: Which of the following characterizes a support vector in soft-margin SVM?

◯ A misclassified training data point.

◯ A correctly classified training data point that is inside the margin.

◯ A correctly classified training data point that is on the margin.

◯ All of the above.

Solution: D. A, B, and C are the three types of support vectors, so all of them characterize a support vetor in soft-margin SVM.

Problem 11: Suppose that you tried to use the hard-margin SVM formulation to find a classification model, but your solver returned no solution. Which of the following changes could you make to your problem formulation that would allow you to find a SVM model?

- ◯ Switch the training labels, so that data points previously labeled as +1 now have label -1, and vice versa.
- ◯ Add slack variables to the problem constraints, i.e., use the soft-margin SVM formulation.
- ◯ Introduce a bias term into your features, i.e., add another feature that takes the value 1 for all training data points.
- ◯ This is a trick question. Hard-margin SVM always yields a solution, so the premise of the question would never happen.

Solution: B. The hard-margin SVM formulation has no feasible solution if the data are not linearly separable. Introducing slack variables allows us to find a solution if the data is not linearly separable.