Exam 1 — Set A

CSCI 567 — Machine Learning — Fall 2026

October 16, 2026

Instructions.

- Time: **2 hours**. Total: **90 points**. Covers Lectures 01–06.
- Closed book. One double-sided handwritten A4/letter sheet of notes is allowed. No calculators, phones, or laptops.
- **Part A (multiple choice, 50 pts):** each question has four options and *one or more* correct answers. Mark every option you believe is correct. Scoring per question: 5 pts if your set of marks is exactly right, 2 pts if it differs from the correct set in exactly one option, 0 otherwise. No justification is needed.
- **Part B (short answer, 30 pts):** answer in a few sentences or a short derivation. Be precise; a correct answer with a wrong reason earns little credit.
- **Part C (problem, 10 pts):** show your work. Partial credit is given for correct intermediate steps.
- Notation is as in lecture: labels $\mathbf{y} \in \{0, 1\}$ for logistic regression and $\mathbf{y} \in \{-1, +1\}$ for the Perceptron and SVM, the margin is $m = \mathbf{y}(\mathbf{w}^\top \mathbf{x} + b)$, $\sigma(z) = 1/(1+e^{-z})$, and $[z]_+ = \max(0, z)$.

Name: _____ USC ID: _____

Part A: multiple choice (10 questions, 5 pts each)

Mark *all* correct options.

A1. A model f is trained on a training set S and then evaluated on a test set T of n examples, drawn i.i.d. from the same distribution $\mathcal{D}$ as S and never used during training. Which statements are true?

(a) The test error of f on T is an unbiased estimate of the true risk $\mathcal{R}(f)$.

(b) The training error of f on S is an unbiased estimate of $\mathcal{R}(f)$.

(c) By Hoeffding's inequality, the width of a confidence interval for $\mathcal{R}(f)$ built from T shrinks like $1/\sqrt{n}$.

(d) If instead f is the best of 1000 candidate models, chosen by picking the lowest error on T, the single-model Hoeffding bound still applies to the chosen model's error on T.

A2. Recht et al. (2019) built a new ImageNet test set following the original collection protocol and re-evaluated many classifiers. Which statements are true?

(a) Every model's accuracy dropped on the new test set.

(b) The drops are explained by sampling noise in the test sets.

(c) The pattern of drops matched the fingerprint of adaptive overfitting: newer, more-tuned models gained less on the new test set.

(d) The evidence points to a distribution gap: the new images are harder for the models, even though a human control showed no comparable difficulty for people.

A3. Let $g : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Which statements are true?

(a) The 0-1 loss, $\mathbb{1}[\mathbf{y}\,\mathbf{w}^\top \mathbf{x} \le 0]$, is a convex function of $\mathbf{w}$.

(b) Any point $\mathbf{w}$ with $\nabla g(\mathbf{w}) = 0$ is a global minimizer of g.

(c) For a sufficiently small constant step size, each gradient descent step on g does not increase g (to first order, a step in direction $-\nabla g$ decreases g).

(d) A stochastic gradient computed from one random training example equals the full gradient at every step.

A4. The Perceptron algorithm (labels $\mathbf{y} \in \{-1, +1\}$, bias folded into $\mathbf{w}$). Which statements are true?

(a) $\mathbf{w}$ is updated only on examples it currently misclassifies.

(b) The algorithm terminates on every dataset.

(c) It is exactly SGD with step size 1 on the loss $\ell(\mathbf{w}; \mathbf{x}, \mathbf{y}) = [-\mathbf{y}\,\mathbf{w}^\top \mathbf{x}]_+$.

(d) On linearly separable data, the separator it returns is the maximum-margin separator.

A5. Polynomial regression of degree M on n noisy samples of a smooth function, using ordinary least squares. Which statements are true?

(a) Increasing M typically decreases bias and increases variance.

(b) Collecting more training data reduces the bias of a fixed model class.

(c) The irreducible noise term in the bias-variance decomposition depends on the model class.

(d) The training MSE is non-increasing in M (for nested classes on the same data).

A6. Ridge (ℓ_2) and Lasso (ℓ_1) regularization, with $\lambda > 0$. Which statements are true?

(a) Ridge has a closed-form solution for every $\lambda > 0$, even when $X^\top X$ is singular.

(b) Lasso can set coefficients exactly to zero.

(c) Ridge is the MAP estimate under a Gaussian prior $\mathbf{w} \sim \mathcal{N}(0, \tau_w^2 I)$ with Gaussian noise.

(d) Increasing λ in Ridge increases $\|\mathbf{w}^\star\|_2$.

A7. Kernels. Which statements are true?

(a) $\kappa(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}')^2$ is a valid (Mercer) kernel.

(b) $\kappa(\mathbf{x}, \mathbf{x}') = -\|\mathbf{x} - \mathbf{x}'\|_2^2$ is a valid kernel.

(c) Training kernel ridge regression costs $O(n^3)$ regardless of the dimension of the feature map.

(d) The representer theorem says the ridge minimizer in feature space lies in the span of the training feature vectors $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)$.

A8. Double descent and implicit regularization for linear models with $\tilde{d}$ features and n training points, trained with squared loss. Which statements are true?

(a) The test error of the minimum-norm interpolant peaks near the interpolation threshold $\tilde{d} = n$.

(b) Gradient descent started from $\mathbf{w} = 0$ on an underdetermined least-squares problem converges to the minimum-norm interpolant.

(c) All interpolating solutions have the same test error.

(d) Gradient descent from a non-zero initialization converges to the same $\mathbf{w}$ regardless of the initialization.

A9. Support vector machines. Which statements are true?

(a) Removing a training point that is not a support vector leaves the hard-margin SVM solution unchanged.

(b) Increasing C in the soft-margin SVM makes margin violations cheaper.

(c) The hinge loss is exactly zero for every point with margin $m \geq 1$.

(d) The logistic loss is exactly zero for points with a large enough positive margin.

A10. Unsupervised learning. Which statements are true?

(a) The first principal component is the eigenvector of the data covariance matrix with the largest eigenvalue.

(b) Lloyd's algorithm for K-means never increases the K-means objective from one iteration to the next.

(c) Lloyd's algorithm always converges to the global minimum of the K-means objective.

(d) K-means is hard EM for a Gaussian mixture with uniform mixing weights and identical spherical covariances.

Part B: short answer (3 questions, 10 pts each)

B1. (10 pts) A colleague trains 200 models with different hyperparameters, evaluates each on the test set, and reports the best of the 200 test accuracies as "our model's accuracy."

(i) **(3 pts)** Explain why the reported number is optimistically biased, and name the complication from Lecture 01 this is an instance of.

(ii) **(3 pts)** State the correct protocol.

(iii) **(4 pts)** For a single fixed model, Hoeffding gives $\mathbb{P}(|\hat{A} - \mu| > \epsilon) \leq 2e^{-2n\epsilon^2}$. Using a union bound over the 200 models, how large must n be so that *every* model's test accuracy is within ϵ of its true accuracy with probability at least $1 - \delta$? Compare with the single-model requirement, and say in one sentence what this means for the cost of testing many models.

B2. (10 pts) Logistic regression on linearly separable data.

(i) **(5 pts)** Prove that logistic loss $\hat{\mathcal{R}}(\mathbf{w}) = \frac{1}{n}\sum_i \log(1 + e^{-\tilde{\mathbf{y}}_i \mathbf{w}^\top \mathbf{x}_i})$ (labels $\tilde{\mathbf{y}}_i \in \{-1, +1\}$) has no finite minimizer when some $\mathbf{w}_0$ separates the data.

(ii) **(5 pts)** Give two different remedies from the course and say, for each, what its minimizer looks like.

B3. (10 pts) Two practical questions about unsupervised learning.

(i) **(5 pts)** Explain why choosing the number of clusters K by minimizing the K-means objective on held-out data does not work, unlike choosing a regularization strength by validation error in supervised learning.

(ii) **(5 pts)** Explain why standardizing the features before PCA matters, with a concrete example of what goes wrong without it.

Part C: problems (1 problem, 10 pts)

C1. (10 pts) A kernel interpolant by hand. Three training points in one dimension: $(x_1, y_1) = (-1, 1)$, $(x_2, y_2) = (0, 0)$, $(x_3, y_3) = (1, 1)$.

(i) **(2 pts)** Show that no model of the form $f(x) = w_1 x + w_0$ fits all three points exactly.

(ii) **(2 pts)** Write down the feature map ϕ for which $\kappa(x, x') = (1 + xx')^2$ equals $\phi(x)^\top \phi(x')$, and compute the 3×3 Gram matrix K on the training points.

(iii) **(4 pts)** Kernel ridge regression predicts $f(x) = \sum_i \alpha_i \kappa(x, x_i)$ with $\alpha = (K + n\lambda I)^{-1}\mathbf{y}$. Take the limit $\lambda \to 0$ and solve for α. (Hint: the data are symmetric about 0, so look for $\alpha_1 = \alpha_3$.)

(iv) **(2 pts)** Write $f(x)$ as an explicit polynomial and check that it interpolates the three points.