Homework 2

Stats 315A, Winter 2026

January 30, 2026

Due date: Wednesday, February 4th at 11:59pm on Canvas.

Submission instructions: Upload a `.pdf` file with your answers, and a separate text file or zip file with your code from question 1.

Question 1 (Conditional validity on CIFAR-100, **20 points**): The CIFAR-100 dataset is a dataset consisting of small images from 100 distinct classes. In this question, you will compare "static" split-conformal methods for predictive inference to "conditional" methods. Assuming you have a predictor $f : \mathcal{X} \to \mathbb{R}^k$, which assigns a score $f_y(x)$ to each class $y \in \{1, \ldots, k\}$ (where $k = 100$ in this case), the static conformal methodology constructs a confidence set of the form

$$\widehat{C}(x) = \{y \in [k] \mid f_y(x) \geq \widehat{\tau}\},$$

where $\widehat{\tau}$ is a threshold, so that $\widehat{C}(x)$ contains classes assigned a high-enough score by the predictive model f. The conditional type method uses

$$\widehat{C}(x) = \left\{y \in [k] \mid f_y(x) \geq \phi(x)^T \widehat{\theta}\right\},$$

where $\widehat{\theta}$ is a fit vector and $\phi : \mathcal{X} \to \mathbb{R}^d$ is a feature function. For each, we assume existence of a validation dataset $Z_{\text{val}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{val}}}$, and for the loss $\ell_\alpha(t) = \alpha\,(t)_+ + (1-\alpha)\,(-t)_+$, choose $\widehat{\tau}$ and $\widehat{\theta}$, respectively, by fitting

$$\widehat{\tau} = \operatorname*{argmin}_{\tau} \frac{1}{n_{\text{val}}} \sum_{x,y \in Z_{\text{val}}} \ell_\alpha(f_y(x) - \tau) \quad \text{and} \quad \widehat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n_{\text{val}}} \sum_{x,y \in Z_{\text{val}}} \ell_\alpha(f_y(x) - \theta^T \phi(x)).$$

Note that these correspond to using the scoring function $s(x, y) = -f_y(x)$ and prediction set

$$C(x) = \{y \in [k] \mid s(x, y) \leq \tau(x)\}$$

for a threshold $\tau(x)$ in the "standard" conformal prediction setup. In this question, for covariates $x \in \mathbb{R}^p$, we will use random feature functions of the form

$$\phi(x) = Wx, \quad \text{where } W \in \mathbb{R}^{d \times p}, \quad W_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$$

to get a sense of coverage.

To get started we provide `cifar_processing.py`. This file contains `process_cifar_into_vectors`, a function that generates data files `cifar100_train_features.npy` and `cifar100_train_labels.npy` (and the similarly named `test` files). These files contain NumPy matrices of the CIFAR-100 train and test datasets processed through a 50 layer Residual network He et al. (2016a,b), yielding data vectors $x \in \mathbb{R}^p$ with $p = 2048$. `cifar_processing.py` also provides methods `load_numpy_into_data` and `split_into_train_and_validation` to help with data processing. Perform the following experiment with this data:

i. Split the non-test data (randomly) into a training dataset of $4 \cdot 10^4$ examples and validation dataset of $n_{\text{val}} = 10^4$ examples, and fit a classifier using multiclass logistic regression (the method `train_prediction_model` from `conformal_multiclass_starter.py` may be helpful). Your classifier should achieve roughly 70+% accuracy on the validation data.

ii. Fit the standard (static) split-conformal predictor $\widehat{\tau}$ using your classifier f using the validation data and $\alpha = .1$.

iii. For a random matrix $W \in \mathbb{R}^{d \times p}$ with $d = 10$ and rows $w_1, \ldots, w_d \in \mathbb{R}^p$, find the best linear predictor of the quantiles $\widehat{\theta}$ as above, with $\alpha = .1$.

iv. Let Z_{test} denote the test data. On this dataset, evaluate the coverage of the resulting confidence sets on "extreme" subsets of the test set defined by inner products $w_i^T x$, that is, the subsets of the test defined by

$$Z_{\text{test},i} := \{(x, y) \in Z_{\text{test}} \mid w_i^T x \geq \text{QUANT}_{.9}(x^T w_i) \;\text{ or }\; w_i^T x \leq \text{QUANT}_{.1}(x^T w_i)\},$$

the smallest and largest 10% of data as defined by $x^T w_i$, for each $i = 1, \ldots, d$. Record both the coverage of $\widehat{C}$ on $Z_{\text{test},i}$ as well as the average confidence set size. Also record the marginal coverage on Z_{test}.

Repeat this experiment 10 times (i.e., over 10 random splits of the training data into train and validation data), and provide a box plot of the coverage and confidence set sizes across the random splits defined by W; provide also the marginal coverage and marginal confidence set sizes on Z_{test}. Describe your observations.

Question 2 (Constructions of conformal confidence sets, **8 points**)**:** Suppose we have set-valued mappings $C_\tau : \mathcal{X} \rightrightarrows \mathcal{Y}$, meaning that $C_\tau(x) \subset \mathcal{Y}$, indexed by $\tau \in \mathbb{R}_+$, where

$$C_\tau(x) \subset C_{\tau+\delta}(x)$$

for all $\delta \geq 0$, where $\lim_{\tau \to \infty} C_\tau(x) = \mathcal{Y}$ (that is, for large enough τ the confidence set $C_\tau(x)$ includes all of $\mathcal{Y}$). Define

$$s(x, y) := \inf \{\tau \in \mathbb{R} \mid y \in C_\tau(x)\}. \tag{1}$$

You are given a sample $(X_i, Y_i)_{i=1}^n \stackrel{\text{iid}}{\sim} P$ of size n and define $S_i = s(X_i, Y_i)$ for each i, then set

$$\widehat{\tau}_n := \text{the } (1 + 1/n)(1 - \alpha) \text{ quantile of } \{S_i\}_{i=1}^n.$$

Let $\widehat{C} = C_{\widehat{\tau}_n}$ be the associated confidence set.

(a) Using the results from class, show that $\widehat{C}$ is a valid $(1 - \alpha)$ prediction set, that is, on a new example (X_{n+1}, Y_{n+1}) from P,

$$P(Y_{n+1} \in \widehat{C}(X_{n+1})) \geq 1 - \alpha.$$

We now explore different constructions of such confidence sets. Each of these will leverage an already constructed predictor f taking inputs in $\mathcal{X}$.

(b) Let ℓ be a loss function and $\ell(f(x), y)$ be the loss for predicting $f(x)$ on response y. Set

$$C_\tau(x) = \{y \in \mathcal{Y} \mid \ell(f(x), y) \leq \tau\}.$$

Give the value $s(x, y)$ the definition (1) yields.

(c) For binary logistic regression, $f : \mathcal{X} \to \mathbb{R}$, $y \in \{\pm 1\}$, and $\ell(f(x), y) = \log(1 + e^{-yf(x)})$. Give the value $s(x, y)$ the definition (1) yields for the confidence set in part (b). For a given threshold τ, when is $C_\tau(x)$ a singleton?

(d) For k-class logistic regression, $f : \mathcal{X} \to \mathbb{R}^k$, $y \in \{1, \ldots, k\}$, and $\ell(f(x), y) = \log(1 + \sum_{l=1}^k e^{f_l(x) - f_y(x)})$. Give the value $s(x, y)$ the definition (1) yields for the confidence set in part (b). For a given threshold τ, when is $C_\tau(x)$ a singleton?

(e) Let $\mathcal{Y} = \mathbb{R}$ (so we have real-valued responses as in regression), and let $l, u : \mathcal{X} \to \mathbb{R}$ model lower and upper quantiles of Y given X, respectively. (That is, we wish to have $Y \in [l(x), u(x)]$ with a given probability.) Let

$$C_\tau(x) = [l(x) - \tau, u(x) + \tau]$$

where $C_\tau(x) = \emptyset$ if $l(x) - \tau > u(x) + \tau$, i.e., the lower end of the interval is greater than the upper. Give the value $s(x, y)$ the definition (1) yields for this confidence set.

Question 3 (Facts about CDFs and quantiles,[1] **6 points**):

In this exercise, we'll walk through a number of basic but important facts about quantiles and cumulative distribution functions (CDFs). Let F be a CDF, of the form

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R},$$

for some real-valued random variable X. Let Q be the corresponding quantile function,

$$Q(t) = \inf\{x : F(x) \geq t\}, \quad t \in [0, 1].$$

(This is often denoted as $Q = F^{-1}$, even when the inverse of F does not exist in the usual sense.) We note that F is always nondecreasing and right-continuous; the latter says, for any x,

$$F(x) = \lim_{y \to x^+} F(y),$$

where $y \to x^+$ means that y approaches x from the right.

(a) Show that for any x and any t,

$$F(x) \geq t \Leftrightarrow Q(t) \leq x.$$

This is sometimes called the *Galois inequality* for the quantile function. Hint: one direction follows from the definition of Q, and the other is a consequence of right-continuity of F.

(b) Use part (a) to show that if $U \sim \text{Unif}(0, 1)$, then $Q(U)$ is distributed according to F (which means it has CDF F).

(c) Use part (a) to show that for any t,

$$F(Q(t)) \geq t,$$

with equality if and only if t is in the range of F.

(d) Use parts (b) and (c) to show that if X is distributed according to F, then $F(X)$ is sub-uniform, which means that for any t,

$$\mathbb{P}(F(X) \leq t) \leq t,$$

with equality if t is in the range of F. Hint: you may start by replacing X with $Q(U)$, for $U \sim \text{Unif}(0, 1)$.

[1] This question is adapted from Tibshirani (2023, homework 4).

(e) Give a concrete worked example to show when equality fails in the result in part (d).

(f) We can always achieve equality in part (d) via auxiliary randomization. Define

$$F^*(x; v) = \lim_{y \to x^-} F(y) + v \cdot \Big(F(x) - \lim_{y \to x^-} F(y) \Big),$$

where $y \to x^-$ means that y approaches x from the left. Show empirically, by revisiting your example in part (e), that for $V \sim \text{Unif}(0, 1)$, independent of X, and for any t,

$$\mathbb{P}(F^*(X; V) \leq t) = t.$$

Question 4 (Holdout sets and adaptive overfitting,[2] **14 points**): The holdout method is a common technique in machine learning to perform model selection. The method holds out a set S of n examples (x_i, y_i) sampled i.i.d. from a distribution D and uses this set to evaluate the performance of a proposed model. Concretely, for a classifier f, one uses the empirical risk $R_S[f]$ on the holdout set S as a proxy for the true model risk $R[f]$. Throughout, assume we have binary labels $y_i \in \{0, 1\}$, and we measure performance using the 0–1 loss.

(a) Fix a classifier f. Show that if

$$n \geq \frac{\log(2/\delta)}{2\varepsilon^2},$$

then with probability $1 - \delta$,

$$\big| R_S[f] - R[f] \big| \leq \varepsilon.$$

Hint: Hoeffding's inequality.

(b) The popular ImageNet ILSVRC and Cifar10 datasets have, respectively, $n = 50{,}000$ and $n = 10{,}000$ images in the validation set. If we set $\delta = 0.05$ (corresponding to a 95% confidence interval), evaluate the bound from part (a) for both ImageNet and Cifar10.

Most machine learning workflows, however, do not evaluate a single classifier on the holdout set and then stop. Instead, after looking at the validation loss, you try to improve it by, for instance, changing the feature set, adding more layers, tweaking the optimization algorithm, etc., and then reevaluate the new model on the same validation set. In the remainder of this problem, we explore the potential pitfalls of adaptively interacting with the holdout set.

Henceforth, suppose our features are binary $x \in \{0, 1\}^d$, and suppose examples (x, y) are drawn from the uniform distribution on $\{0, 1\}^d \times \{0, 1\}$. Consider the following procedure:

(c) Compute R_S for single-feature classifiers $h_i(x) = x_i$ for $i = 1, \ldots, d$.

(d) Say a feature i is *informative* if

$$R_S[h_i] \leq \frac{1}{2} - \sqrt{\frac{1}{n}}.$$

Let I denote the set of informative classifiers h_i.

[2] This question is adapted from Hardt and Recht (2025).

(e) Construct a classifier $\tilde{f}$ consisting of a majority vote of the informative classifiers:

$$\tilde{f}(x) = \begin{cases} 1 & \text{if } \sum_{i \in I} h_i(x) \geq \frac{|I|}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

This is a fairly natural procedure: we first attempt to predict y using a single feature, and then ensemble the classifiers that seem to give predictive power. However, we will show the estimated risk $R_S[\tilde{f}]$ can be arbitrarily far from the true risk $R[\tilde{f}]$ when d is large.

(f) First, prove

$$R[\tilde{f}] = \frac{1}{2}.$$

This means $\tilde{f}$ is no better than random guessing on new examples. *Hint: You can in fact prove that $R[f] = \frac{1}{2}$ for any arbitrary classifier f.*

(g) (Bonus) Show the expected empirical risk of $\tilde{f}$ shrinks exponentially fast with the number of informative features $|I|$. Prove

$$\mathbb{E}_S\Big[\, R_S[\tilde{f}] \mid |I| = k \,\Big] \leq \exp\left(-\frac{2k}{n}\right).$$

Even if you don't solve the bonus exercise, we'll use the conclusion in the subsequent parts. *Hint: Use the fact that the coordinates are independent, so $\mathbf{1}[x_i = y]$ and $\mathbf{1}[x_j = y]$ are independent for $i \neq j$, along with the observation $\tilde{f}(x) \neq y$ iff $\sum_{i \in I} \mathbf{1}[x_i = y] < |I|/2$.*

(h) The remainder of the problem is devoted to showing $|I|$ is large with high probability. Prove each coordinate i is informative with constant probability, i.e. show

$$\mathbb{P}\{i \in I\} \geq c$$

for some constant $c > 0$. Hint: First, argue $R_S[h_i]$ follows a rescaled binomial distribution, and $i \in I$ if the binomial deviates from its mean by 2 standard deviations. Then, show this event occurs with constant probability by approximating the binomial to a normal distribution. You don't need to be rigorous with the approximation.

(i) Prove $\mathbb{E}[|I|] \geq cd$.

(j) Prove with probability $1 - \delta$, the number of informative features satisfies

$$|I| \geq \frac{cd}{2}$$

for

$$d \geq \frac{2\log(1/\delta)}{c^2}.$$

Hint: Use that each coordinate is independent, so $\mathbf{1}[x_i \neq y]$ and $\mathbf{1}[x_j \neq y]$ are independent for $i \neq j$, and then apply Hoeffding's inequality.

(k) Put parts (c)–(g) together to prove the following: there exists a constant α such that if $d \geq \alpha n$, then with probability at least 3/4,

$$R[\tilde{f}] - R_S[\tilde{f}] \geq 0.49.$$

Question 5 (Time and collaboration, **2 points**)**:**

(a) How many hours in total did you spend on this assignment? (This will help to calibrate future assignments.)

(b) With whom (if anyone) did you collaborate on this assignment?

References

Moritz Hardt and Benjamin Recht. Patterns, predictions, and actions: A story about machine learning. https://mlstory.org/, 2025. Online textbook.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 2016b.

Ryan Tibshirani. Advanced topics in statistical learning (stat 241b / cs 281b), spring 2023. https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/, 2023. Course website, University of California, Berkeley.