Evaluating Machine Learning Techniques

Fabio G. Cozman - fgcozman@usp.br

September 24, 2019

Training / validation / testing datasets

- A classifier is produced using the training set.
- Hyperparameters may be set using a validation set.
- Then the error rate is estimated using the testing set.

A few points:

- More training data is better to get better classifier, more testing is better to estimate error rate.

- Never use the testing set to choose any aspect of a classifier!
- Not enough data: use cross-validation!
  - Note: with cross-validation we are not evaluating a single classifier; rather, we are evaluating the learning method.

Cross-validation

- If not enough data to create a *holdout* to test, at least *cross-validation* must be used.
- Idea: separate a fraction of the data for testing, then repeat over the whole database.
- Five-fold or ten-fold cross-validation are very common.

Other cross-validation techniques

- *Leave-one-out cross validation (LOOCV)* is also common but it demands more computation (number of folds = number of data points).

- *Stratified cross-validation*: divide folds but respect proportion of labels within each fold.

Other metrics

- Accuracy / error rate is not the only metric.
- One can build a *confusion matrix*.
- Or one can measure *precision* and *recall*.
- Or one can build a *ROC* curve.
- Or one can compute the *AUC*.

Confusion matrix

	Actual label		
	Car	Bus	Bike
Car	20	2	0
Bus	3	33	0
Bike	1	1	18

That is, one car was classified as a bike, while 2 buses were classified as cars, and so on.

Confusion matrix: two labels

	Actual label: The label	Actual label: Not the label
The label	true positives (TP)	false positives (FP)
Not the label	false negatives (FN)	true negatives (TN)

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} = \frac{TP+TN}{N} = \frac{\text{true}}{\text{all}}.$$

Problem with accuracy

- Class imbalance: one label is much more common than other.
  - Suppose 95% of images do not contain camels; a camel detector that always outputs NO has accuracy 0.95.
- Also, some errors are worse than others.
  - Medical domain: a false positive (detecting a disease when there is none) is awful but a false negative (failing to detect a disease) may be deadly.

Precision and Recall

	Actual label: The label	Actual label: Not the label
The label	true positives (TP)	false positives (FP)
Not the label	false negatives (FN)	true negatives (TN)

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{\text{true positive}}{\text{get the label}}.$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{\text{true positive}}{\text{have the label}}.$$

More on Precision and Recall

- Intuition: precision is the probability that a (randomly) selected input is actually relevant.
- Intuition: recall is the probability that a (randomly) relevant input is actually selected.

More on Precision and Recall

- Intuition: precision is the probability that a (randomly) selected input is actually relevant.
- Intuition: recall is the probability that a (randomly) relevant input is actually selected.

- Recall is also called *true positive rate* or *sensitivity*.
- There is also the *true negative rate* or *specificity*.

True negative rate

	Actual label: The label	Actual label: Not the label
The label	true positives (TP)	false positives (FP)
Not the label	false negatives (FN)	true negatives (TN)

$$\text{True negative rate} = \frac{TN}{TN+FP} = \frac{\text{true negative}}{\text{do not have the label}}.$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{\text{true positive}}{\text{have the label}}.$$

F_1-score

- Precision $= p$; Recall $= r$.
- F_1-score is the harmonic mean:

$$\left(\frac{p^{-1} + r^{-1}}{2}\right)^{-1} = \frac{2}{1/p + 1/r} = 2\frac{pr}{p + r}.$$

- When $p = r = 1$, then F_1-score $= 1$.
- But if p or r is zero, then F_1-score $= 0$.

Generalizing the F_1 score

- There is F_β, when one attaches more importance to recall than to precision as β grows.
- Definition

$$
\begin{aligned}
F_\beta &= \left(\frac{p^{-1} + \beta^2 r^{-1}}{1+\beta^2}\right)^{-1} = \frac{(1+\beta^2)pr}{\beta^2 p + r} \\
&= \frac{1}{\frac{\alpha}{p} + \frac{1-\alpha}{r}} \quad \text{(where } \alpha = 1/(1+\beta^2)\text{)}.
\end{aligned}
$$

True negative / false positive rate

	Actual label: The label	Actual label: Not the label
The label	true positives (TP)	false positives (FP)
Not the label	false negatives (FN)	true negatives (TN)

$$\text{True negative rate} = \frac{TN}{TN+FP} = \frac{\text{true negative}}{\text{do not have the label}}.$$

$$\text{False positive rate} = \frac{FP}{FP+TN} = \frac{\text{false positive}}{\text{do not have the label}}.$$

The ROC curve

- Suppose we have a classifier that
  - produces $h(\mathbf{x})$ for a given observed $\mathbf{x}$;
  - then $\hat{Y} = 1$ if $h(\mathbf{x}) > \alpha$ and $\hat{Y} = 0$ otherwise.

- The Receiver Operating Characteristic (ROC) curve captures the influence of α.

Understanding α for plug-ins...

- Start with Bayes classifier: minimize

$$\mathbb{P}(g(\mathbf{X}) \neq Y) = \sum_{\mathbf{x}} \mathbb{P}(g(\mathbf{x}) \neq Y | \mathbf{X} = \mathbf{x}) \, \mathbb{P}(\mathbf{X} = \mathbf{x}) \, .$$

Understanding α for plug-ins...

- Start with Bayes classifier: minimize

$$\mathbb{P}(g(\mathbf{X}) \neq Y) = \sum_{\mathbf{x}} \mathbb{P}(g(\mathbf{x}) \neq Y | \mathbf{X} = \mathbf{x}) \, \mathbb{P}(\mathbf{X} = \mathbf{x}) \, .$$

- So, minimize $\mathbb{P}(g(\mathbf{x}) \neq Y | \mathbf{X} = \mathbf{x})$ for each $\mathbf{x}$.

Understanding α for plug-ins...

- Start with Bayes classifier: minimize

$$\mathbb{P}(g(\mathbf{X}) \neq Y) = \sum_{\mathbf{x}} \mathbb{P}(g(\mathbf{x}) \neq Y | \mathbf{X} = \mathbf{x}) \, \mathbb{P}(\mathbf{X} = \mathbf{x}) \, .$$

- So, minimize $\mathbb{P}(g(\mathbf{x}) \neq Y | \mathbf{X} = \mathbf{x})$ for each $\mathbf{x}$.
- Hence the Bayes classifier is:

$$\hat{Y} = g(\mathbf{x}) = \begin{cases} 1 & \text{if } \frac{\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=0|\mathbf{X}=\mathbf{x})} > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now, missclassification costs

- Assume the cost $c(\hat{Y}, Y)$:

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	0	a
$\hat{Y} = 0$	b	0

- How to minimize $\mathbb{E}\Big[c(\hat{Y}, Y)\Big]$?

Minimizing the expected loss

- Minimize:

$$\mathbb{E}\Big[c(\hat{Y}, Y)\Big] = \sum_{\mathbf{x}} \begin{pmatrix} b\mathbb{P}(Y = 1|\mathbf{x}) & \text{if } g(\mathbf{x}) = 0, \\ a\mathbb{P}(Y = 0|\mathbf{x}) & \text{if } g(\mathbf{x}) = 1 \end{pmatrix} \mathbb{P}(\mathbf{X} = \mathbf{x}) \,.$$

Minimizing the expected loss

- Minimize:

$$\mathbb{E}\Big[c(\hat{Y}, Y)\Big] = \sum_{\mathbf{x}} \begin{pmatrix} b\mathbb{P}(Y=1|\mathbf{x}) & \text{if } g(\mathbf{x})=0, \\ a\mathbb{P}(Y=0|\mathbf{x}) & \text{if } g(\mathbf{x})=1 \end{pmatrix} \mathbb{P}(\mathbf{X}=\mathbf{x}).$$

- So:

$$\hat{Y} = g(\mathbf{x}) = \begin{cases} 1 & \text{if } \frac{\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=0|\mathbf{X}=\mathbf{x})} > \frac{a}{b}, \\ 0 & \text{otherwise.} \end{cases}$$

So, for plug-ins:

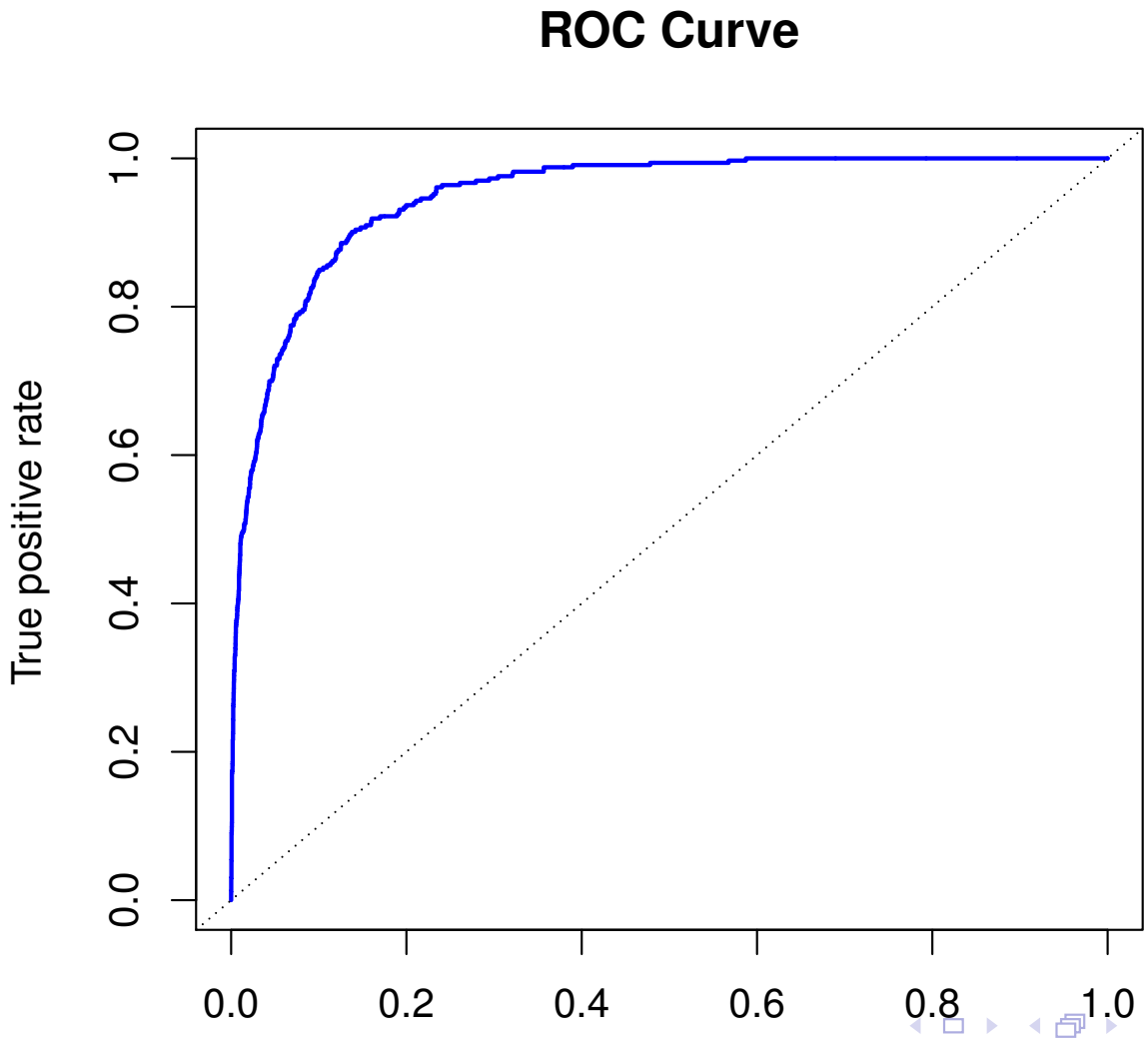
- Classifier
  - produces $h(\mathbf{x})$ for a given observed $\mathbf{x}$;
  - then $\hat{Y} = 1$ if $h(\mathbf{x}) > \alpha$ and $\hat{Y} = 0$ otherwise.
- $h(\mathbf{x}) = \frac{\mathbb{P}(Y=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y=0|\mathbf{X}=\mathbf{x})}$.
- $\alpha = \frac{a}{b}$.

Intuition behind ROC curve

- With a lot of data:
  - the true positive rate is "close" to $\mathbb{P}(h(\mathbf{X}) > \alpha | Y = 1)$;
  - the false positive rate is "close" to $\mathbb{P}(h(\mathbf{X}) > \alpha | Y = 0)$.
- The ROC curve captures the change in these probabilities as α varies.

Drawing a ROC curve

- Idea: plot true-positive rates for false-positive rates.



Properties of ROC curves

- $(1, 1)$ is a point for "very small" α.
- $(0, 0)$ is a point for "very large" α.
- $(0, 1)$ is really the best possible point.
- Line between $(0, 0)$ and $(1, 1)$ the the "worst" (output is independent of Y!).
- Best curve: from $(0, 0)$ to $(0, 1)$ to $(1, 1)$.

Using the ROC curve

- One can select the best α by finding a point of interest.
- One can compare the curves for various classifiers.
- One can see how close to $(0, 1)$ one can get.

Area Under the Curve

- For a classifier, one can compute the area under the ROC curve: this is the AUC.
  - Best AUC: equal to one.
  - An AUC of $1/2$ is *very bad*...
  - Fact: the AUC estimates $\mathbb{P}(h(\mathbf{x}') > h(\mathbf{x}'') | Y' = 1, Y'' = 0)$.

Ablation study

- Often a system based on machine learning has many features, many components.
- Idea: remove one part at a time, check accuracy, F_1 score, ROC curve, ...
- This is called an *ablation* study.

Other tasks

- Confidence intervals (for error rates).
- Comparing classifiers:
  - Is classifier A better than classifier B?
  - Compute expected performance of A and B?
  - Compute accuracy, etc, for both. Are they "different enough"?

Frequentist statistics

- Suppose prior cannot be specified. What to do?
- For estimation, maximum likelihood (others: least-squares, method of moments, invariance principles...).
- Confidence intervals are widely used.
- For decisions, hypothesis testing, p-values, analysis of variance...

Maximum likelihood

- Take:
$$\hat{\theta} = \arg\max p(x|\theta)\,,$$
- This is the *maximum likelihood* estimator.
- Function $p(x|\theta)$ is the likelihood function, denoted by $L(\theta)$.

An example I

- Take $p(x|\theta) = (2\pi)^{-1/2} \exp(-(x-\theta)^2/2)$.
- The maximum likelihood estimate is x.
- Suppose we have a sequence of independent identically distributed measurements $X_1, \ldots, X_N$, then $p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$.
- Consequently, $p(x_1, \ldots, x_n|\theta)$ is proportional to $\exp(-\sum_{i=1}^{N}(x_i-\theta)^2/2)$.

An example II

- The maximum likelihood estimate is:

$$
\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} \exp(-\sum_{i=1}^{N}(x_i-\theta)^2/2) \\
&= \arg\min_{\theta} \sum_{i=1}^{N} x_i^2 - 2x_i\theta + \theta^2 \\
&= \arg\min_{\theta} \sum_{i=1}^{N} x_i^2 - 2\theta\sum_{i=1}^{N} x_i + \sum_{i=1}^{N}\theta^2 \\
&= \arg\min_{\theta} \sum_{i=1}^{N} x_i^2 - (1/N)(\sum_{i=1}^{N} x_i)^2 + (1/N)(\sum_{i=1}^{N} x_i)^2 - 2\theta\sum_{i=1}^{N} x_i + N\theta^2 \\
&= \arg\min_{\theta} N(\sum_{i=1}^{N} x_i^2/N - ((\sum_{i=1}^{N} x_i)/N)^2) + N(\theta - \sum_{i=1}^{N} x_i/N)^2.
\end{aligned}
$$

- We minimize this last expression by taking $\hat{\theta} = \sum_i x_i/N$; this is the maximum likelihood estimator.

Confidence intervals

- A $(1-\alpha)$ confidence interval for θ is an interval $[a(X_1,\ldots,X_N), b(X_1,\ldots,X_N)]$ such that

$$\mathbb{P}(\theta \in [a,b]) \geq 1-\alpha$$

  for all θ.
- Note: the interval is random, not θ!
  - So $1-\alpha$ is not probability of θ given data; it is the probability of making the right decision about intervals $1-\alpha$ of the time...

An example: the Gaussian case

- Suppose X_i is Gaussian with mean θ and variance 1.
  - Average is estimator $\hat{\theta} = \sum_i X_i / N$.

An example: the Gaussian case

- Suppose X_i is Gaussian with mean θ and variance 1.
  - Average is estimator $\hat{\theta} = \sum_i X_i / N$.
- Consider interval with confidence $1 - \alpha$:

$$[a, b] = \left[\hat{\theta} - \frac{z}{\sqrt{N}}, \hat{\theta} + \frac{z}{\sqrt{N}} \right],$$

An example: the Gaussian case

- Suppose X_i is Gaussian with mean θ and variance 1.
  - Average is estimator $\hat{\theta} = \sum_i X_i / N$.
- Consider interval with confidence $1 - \alpha$:

$$[a, b] = \left[\hat{\theta} - \frac{z}{\sqrt{N}}, \hat{\theta} + \frac{z}{\sqrt{N}}\right],$$

where z is the value such that

$$\mathbb{P}(-z \leq Z \leq z) = 1 - \alpha$$

for Z Gaussian with mean 0, variance 1.

Example: The Gaussian case

- Then

$$\begin{aligned}
\mathbb{P}(\theta \in [a,b]) &= \mathbb{P}\left(\hat{\theta} - \frac{z}{\sqrt{N}} \leq \theta \leq \hat{\theta} + \frac{z}{\sqrt{N}}\right) \\
&= \mathbb{P}\left(\theta - \frac{z}{\sqrt{N}} \leq \hat{\theta} \leq \theta + \frac{z}{\sqrt{N}}\right) \\
&= \mathbb{P}\left(-z \leq \frac{\hat{\theta} - \theta}{1/\sqrt{N}} \leq z\right) \\
&= 1 - \alpha,
\end{aligned}$$

  regardless of θ (as average is Gaussian with mean θ and variance $1/N$).
- Thus $[a, b]$ is a confidence interval with confidence $1 - \alpha$.

Example: getting z

Distribuição Normal P(0≤Z<z0)										
z0	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Another example: error rate

- Suppose empirical error rate is $\hat{e}_r$ from testing data (M observations).

Another example: error rate

- Suppose empirical error rate is $\hat{e}_r$ from testing data (M observations).
- Then the following interval

$$\left[\hat{e}_r - z\sqrt{\frac{\hat{e}_r(1-\hat{e}_r)}{M}}, \hat{e}_r + z\sqrt{\frac{\hat{e}_r(1-\hat{e}_R)}{M}}\right]$$

includes error rate with probability $1-\alpha$, for some z (for instance, for $\alpha = 0.05$, if $M > 30$ then $z \approx 1.96$).

Another example: error rate

- Suppose empirical error rate is $\hat{e}_r$ from testing data (M observations).
- Then the following interval

$$\left[\hat{e}_r - z\sqrt{\frac{\hat{e}_r(1-\hat{e}_r)}{M}}, \hat{e}_r + z\sqrt{\frac{\hat{e}_r(1-\hat{e}_R)}{M}}\right]$$

  includes error rate with probability $1-\alpha$, for some z (for instance, for $\alpha = 0.05$, if $M > 30$ then $z \approx 1.96$).
- For better estimates: use bootstrap (not covered here...).

Yet another example...

- θ is real number, $Y_i = X_i + \theta$ where X_i are independent, binary with

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2.$$

- Observe Y_1, Y_2, take

$$[a, b] = \begin{cases} [y_1 - 1, y_1 - 1] & \text{if } y_1 = y_2 \\ [(y_1 + y_2)/2, (y_1 + y_2)/2] & \text{if } y_1 \neq y_2. \end{cases}$$

- Now, $\mathbb{P}(\theta \in [a, b]) = 3/4$ for all θ!
- But suppose $y_1 = 15$, $y_2 = 17$.
  1. Then $[a, b] = [16, 16]$ is 75% confidence interval.
  2. And we are *sure* that $\theta = 16$!!

Hypothesis testing

- Start with *null hypothesis* and *alternative hypothesis*

$$H_0 : \theta \in \Theta_0; \qquad H_1 : \theta \in \Theta_1.$$

- Collect data, check whether null hypothesis is rejected: rejection when data falls within a *rejection region* R.
- Note: assume innocence (retain H_0) until proved guilty (data into rejection region).

Errors

- Errors:
  - Type I error: reject H_0 when it is true.
  - Type II error: do not reject H_0 when it is false.

- Main idea is to choose rejection region to "control" Type I error.
  - Level is α when probability of Type I error is less than or equal to α whenever H_0 is true.

Example

- X_i are $N(\theta, \sigma)$ with known σ.
- Test

$$\begin{cases} H_0: & \theta \leq 0, \\ H_1: & \theta > 0. \end{cases}$$

- Rejection region:

$$R = \left\{ x_1, \ldots, x_n : \sum_i x_i / N > c \right\}.$$

- How to choose c?
  - Select c so that $\mathbb{P}(X \in R|\theta) \leq \alpha$ for all θ for H_0.

Difficulties

1. A test with the smallest Error Type II for each size α is a *most powerful* test.
2. These tests are very hard to find.
3. They may not even exist for a given problem.

P-values

- Take a test: reject H_0 iff $T(x_1, \ldots, x_N) \geq c$.
- The p-value is
  - the smallest level α at which we can reject H_0.
  - OR: the (maximum) probability that $T(X_1, \ldots, X_N)$ is even larger than $T(x_1, \ldots, x_N)$ when H_0 is true.

P-values

- Take a test: reject H_0 iff $T(x_1, \ldots, x_N) \geq c$.
- The p-value is
  - the smallest level α at which we can reject H_0.
  - OR: the (maximum) probability that $T(X_1, \ldots, X_N)$ is even larger than $T(x_1, \ldots, x_N)$ when H_0 is true.
- Evidence against H_0: the smaller the p-value, the stronger the evidence against H_0.
- Typically,
  - < 0.01 is very strong evidence against H_0;
  - > 0.1 is little evidence against H_0.

Comparing classifiers: Paired test

- Suppose we have two classifiers and a test dataset.
- Run k-fold cross-validation, applying both classifiers in each fold, so as to have k pairs $(a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)$.
- Assume the distribution of accuracies is Gaussian with identical unkown variance (very strong assumptions!!).
- Hypothesis H_0: there is no difference.

Wald test

- Wald test (large sample) with level α: reject H_0 if $|W| > z_{\alpha/2}$ where

$$W = \frac{\sum_{i=1}^{N}(a_i - b_i)/N}{\sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}\left(a_i - b_i - \frac{\sum_{j=1}^{N}(a_j - b_j)}{N}\right)^2}}.$$

Wald test

- Wald test (large sample) with level α: reject H_0 if $|W| > z_{\alpha/2}$ where

$$W = \frac{\sum_{i=1}^{N}(a_i - b_i)/N}{\sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}\left(a_i - b_i - \frac{\sum_{j=1}^{N}(a_j - b_j)}{N}\right)^2}}.$$

and $z_{\alpha/2}$ is a value such that

$$\mathbb{P}\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha$$

for Z a Gaussian variable with mean 0 and variance 1.

A confidence interval

- Because W is asymptotically Gaussian, we have a confidence interval for the difference between classifiers:

$$\left[\frac{\sum_{i=1}^{N}(a_i - b_i)}{N} - z_{\alpha/2}s, \frac{\sum_{i=1}^{N}(a_i - b_i)}{N} + z_{\alpha/2}s\right],$$

where

$$s = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}\left(a_i - b_i - \frac{\sum_{j=1}^{N}(a_j - b_j)}{N}\right)^2}.$$

Paired t-test

- For smaller samples, assumption of Gaussian distribution for W fails.
- In fact, W is distributed according to t-distribution with $N-1$ degrees of freedom.
- Test: reject H_0 if $|W| > t_{N-1,\alpha/2}$.

Some problems with p-value

- p-value is related to $\mathbb{P}(D|H_0)$; this is not $\mathbb{P}(H_0|D)$!
- the larger the dataset, the more we are likely to reject H_0 no matter what (the null is *always* false...)!

Extra: some bits of frequentist statistics

1. Frequentist measures for estimators.
2. Properties of maximum likelihood estimation.

Just for reference: this material will not be covered!

Frequentist measures: Unbiasedness

- An estimator $\hat{\theta}$ is *unbiased* if
$$\mathbb{E}_\theta\left[\hat{\theta}\right] = \theta$$
for every θ.
- $b_\theta(\hat{\theta}) = \theta - \mathbb{E}_\theta\left[\hat{\theta}\right]$ is the *bias* of the estimator $\hat{\theta}$.
- Reasonable to expect that "good" estimator are unbiased.
- But, there are problems that have no unbiased estimator, and there are problems where unbiased estimators do not satisfy other "reasonable" criteria (for example, maximize the likelihood function).

Frequentist measures: Consistency

- An estimator is consistent if

$$\mathbb{P}\left(\lim_{n\to\infty} \hat{\theta} = \theta\right) = 1.$$

- That is, the estimator certainly takes the "correct" value as we collect an infinite amount of data.

Frequentist measures: Efficiency

- Two unbiased estimators can be compared by their variance.
- We say that $\hat{\theta}_1$ is *more efficient* than $\hat{\theta}_2$ if the variance of $\hat{\theta}_1$ is smaller than the variance of $\hat{\theta}_2$ for all θ.
- This is a "relative" measure of efficiency.

The Cramer-Rao Lower Bound

- For any estimator, the Cramer-Rao lower bound is

$$V_\theta[\hat{\theta}] \geq \frac{(1 - \dot{b}_\theta(\hat{\theta}))^2}{nI(\theta)},$$

where $I(\theta)$ is the Fisher information matrix:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln p(X|\theta)}{\partial \theta^2}\right] = \mathbb{E}\left[\left(\frac{\partial \ln p(X|\theta)}{\partial \theta}\right)^2\right].$$

- We then say that an unbiased estimator $\hat{\theta}$ is *efficient* if

$$V_\theta[\hat{\theta}] = \frac{1}{nI(\theta)}.$$

Frequentist measures: MSE

- A popular measure of estimator quality is the Mean Square Error:

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}_\theta\left[(\hat{\theta}-\theta)^2\right] \\ &= \mathbb{E}_\theta\left[\hat{\theta}^2\right] - 2\theta\mathbb{E}_\theta\left[\hat{\theta}\right] + \theta^2 \pm \mathbb{E}_\theta\left[\hat{\theta}\right]^2 \\ &= \left(\mathbb{E}_\theta\left[\hat{\theta}^2\right] - \mathbb{E}_\theta\left[\hat{\theta}\right]^2\right) - \left(\mathbb{E}_\theta\left[\hat{\theta}\right] - \theta\right)^2 \\ &= V_\theta[\hat{\theta}] + b_\theta^2(\hat{\theta}) \end{aligned}$$

- This is, again, the "bias-variance" decomposition.

Properties of maximum likelihood

- If a maximum likelihood estimator is unbiased, then it is the estimator with minimum variance.
- Adopting some minimal regularity conditions, all maximum likelihood estimators are consistent.
- Adopting some minimal regularity conditions, maximum likelihood estimators are asymptotically efficient, unbiased and normal with

$$\sqrt{nI(\theta)}\left(\hat{\theta}-\theta\right) \sim N(0,1).$$

That is, the asymptotic variance is $(nI(\theta))^{-1}$.

A final note

Some of the figures in this presentation are taken from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.