Prediction

Vicki Hertzberg

31 March 2021

Section 1

Prediction

Subsection 1

Models

Generalized Linear Models

Many variables have distributions that are consistent with an exponential family.

This is the form $f(y|\theta) = exp(\eta(\theta)T(y) - A(\theta) + B(y))$.

Examples are

1. Normal: $f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

Generalized Linear Models

Many variables have distributions that are consistent with an exponential family.

This is the form $f(y|\theta) = exp(\eta(\theta)T(y) - A(\theta) + B(y))$.

Examples are

1. Normal: $f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$
2. Bernoulli: $f(y|\pi) = y^{\pi}(1-y)^{1-\pi}$

Generalized Linear Models

Many variables have distributions that are consistent with an exponential family.

This is the form $f(y|\theta) = exp(\eta(\theta)T(y) - A(\theta) + B(y))$.

Examples are

1. Normal: $f(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$
2. Bernoulli: $f(y|\pi) = y^{\pi}(1-y)^{1-\pi}$
3. Exponential: $f(y|\lambda) = \lambda e^{\lambda y}$ for $y \geq 0$.

Generalized Linear Models

Note that for these distributions we have the following:

1. Normal: $E(Y) = \mu$

Generalized Linear Models

Note that for these distributions we have the following:

1. Normal: $E(Y) = \mu$
2. Bernoulli: $E(Y) = \pi$

Generalized Linear Models

Note that for these distributions we have the following:

1. Normal: $E(Y) = \mu$
2. Bernoulli: $E(Y) = \pi$
3. Exponential: $E(Y) = \lambda^{-1}$

Generalized Linear Model

For the exponential family we can say $E(Y) = A(\theta)$

Consider the concept of simple linear regression, which we can write as $E(Y) = \boldsymbol{X}\beta$

If we extend the concept of regression to this, we can create a generalized linear model, writing $E(Y) = A(\boldsymbol{X}\beta)$.

Generalized Linear Model

The generalized linear model consists of three parts:

1. A probability function from the exponential family

Generalized Linear Model

The generalized linear model consists of three parts:

1. A probability function from the exponential family
2. A linear predictor, $\eta = \boldsymbol{X}\beta$

Generalized Linear Model

The generalized linear model consists of three parts:

1. A probability function from the exponential family
2. A linear predictor, $\eta = \boldsymbol{X}\beta$
3. A link function, $g(\cdot)$, such that $E(y) = g^{-1}(\boldsymbol{X}\beta)$ (or $g(E(y)) = \boldsymbol{X}\beta$)

Generalized Linear Model

For the simple linear regression case we have:

1. The probability function is a normal distribution

Generalized Linear Model

For the simple linear regression case we have:

1. The probability function is a normal distribution
2. The link function is $g(E(y)) = \boldsymbol{X}\beta$

Generalized Linear Model

For *logistic regression* we have:

1. The probability function is a Bernoulli distribution

Generalized Linear Model

For *logistic regression* we have:

1. The probability function is a Bernoulli distribution
2. The link function is $g(E(y)) = \frac{e^{\boldsymbol{X}\beta}}{1+e^{\boldsymbol{X}\beta}}$

Generalized Linear Model

For *time to event* we have

1. The probability function is an exponential distribution

Generalized Linear Model

For *time to event* we have

1. The probability function is an exponential distribution
2. The link function is $g(E(y)) = (\boldsymbol{X}\beta)^{-1}$

Subsection 2

Explaining v Predicting

What do we mean by Explaining?

We postulate the $\mathcal{X}$ causes $\mathcal{Y}$ via a function $\mathcal{F}$. We operationalize this as a statistical model, $E(y) = f(\boldsymbol{X})$, where $f \in \mathcal{F}$

Usually as scientists we want to tease out *causal* relationships.

In an explanatory model, we want to determine how a set of independent variables, $(\boldsymbol{X})$, relates to a dependent outcome, Y.

Goal: to have adequate power and minimal bias in order to match f to $\mathcal{F}$ as closely as possible.

What do we mean by Predicting?

$E(y) = f(\boldsymbol{X})$, $f \in \mathcal{F}$

Statisticians are poor at *prediction*, in general.

In a predictive model, we want to determine how well the relationships between a set of independent variables, ($\boldsymbol{X}$), and a dependent outcome, Y, can be used to predict a new or future observation.

Goal: to use f to generate new Y values.

What difference does it make?

Explaining:

1. f represents the causal relationship between $\boldsymbol{X}$ and Y

What difference does it make?

Explaining:

1. f represents the causal relationship between $\boldsymbol{X}$ and Y
2. f is estimated to test a causal hypothesis

What difference does it make?

Explaining:

1. f represents the causal relationship between $\boldsymbol{X}$ and Y
2. f is estimated to test a causal hypothesis
3. Modeling is retrospective, always dealing with data that have been collected, and f is used to test a pre-existing hypothesis or set of hypotheses

What difference does it make?

Predicting

1. f captures the *association* between $\boldsymbol{X}$ and Y

What difference does it make?

Predicting

1. f captures the *association* between $\boldsymbol{X}$ and Y
2. f is constructed from the data and direct interpretability of the relations between $\boldsymbol{X}$ and Y is not necessary or required.

What difference does it make?

Predicting

1. f captures the *association* between $\boldsymbol{X}$ and Y
2. f is constructed from the data and direct interpretability of the relations between $\boldsymbol{X}$ and Y is not necessary or required.
3. forward-looking, with f being used to predict a new observation Y

What difference does it make?

Finally, consider the expected predicted error for a new observation at value x with a quadratic loss function:

$$EPE = E(Y - \hat{f}(x))^2 \tag{1}$$

$$= E(Y - f(x))^2 - [E(\hat{f}(x)) - f(x)]^2 + E(\hat{f}(x) - E(\hat{f}(x)))^2 \tag{2}$$

$$= Var(Y) + Bias^2 + Var(\hat{f}(x)) \tag{3}$$

Bias results when you mis-specify f. Estimation variance (3rd term) results from using a sample to estimate f. The first term is error that results even if the model is correctly specified and accurately estimated.

What difference does it make?

Explaining: focus is on minimizing bias

Prediction: focus is on minimizing the combination of bias and estimate variance.

Thus sometimes the "wrong" model can predict better than the correct one.

Section 2

Metrics

Metrics for explanatory models

Focus: estimated strength of relationship

1. R^2

Metrics for explanatory models

Focus: estimated strength of relationship

1. R^2
2. Statistical significance with overall F-type statistics

Metrics for predictive models

Focus: predictive accuracy / predictive power (performance of $\hat{f}$ on new data)

Validate model with a *new* dataset, typically by performing cross-validation.

1. Divide data into 10 parts.

Metrics for predictive models

Focus: predictive accuracy / predictive power (performance of $\hat{f}$ on new data)

Validate model with a *new* dataset, typically by performing cross-validation.

1. Divide data into 10 parts.
2. Estimate f from parts 1 - 9, then apply to part 10, how well do you do?

Metrics for predictive models

Focus: predictive accuracy / predictive power (performance of $\hat{f}$ on new data)

Validate model with a *new* dataset, typically by performing cross-validation.

1. Divide data into 10 parts.
2. Estimate f from parts 1 - 9, then apply to part 10, how well do you do?
3. Repeat by estimating from parts 1-8 and 10, then apply to part 9 and assess.

Metrics for predictive models

Focus: predictive accuracy / predictive power (performance of $\hat{f}$ on new data)

Validate model with a *new* dataset, typically by performing cross-validation.

1. Divide data into 10 parts.
2. Estimate f from parts 1 - 9, then apply to part 10, how well do you do?
3. Repeat by estimating from parts 1-8 and 10, then apply to part 9 and assess.
4. Lather, rinse, and repeat until you have done this 10 times.

Metrics for predictive models

Focus: predictive accuracy / predictive power (performance of $\hat{f}$ on new data)

Validate model with a *new* dataset, typically by performing cross-validation.

1. Divide data into 10 parts.
2. Estimate f from parts 1 - 9, then apply to part 10, how well do you do?
3. Repeat by estimating from parts 1-8 and 10, then apply to part 9 and assess.
4. Lather, rinse, and repeat until you have done this 10 times.
5. Metrics include area under the Receiver Operating Characteristic curve (ROC) (typically called AUC), confusion matrix, calibration, discrimination ability, Akaike Information Criterion (AIC)

Section 3

Practicalities

Subsection 1

For better clinical prediction models

Seven Steps for Development

Step 1: What is the research question and what do the data look like?

Step 2: Code predictors

Step 3: Model specification

Step 4: Model estimation

Step 5: Evaluate model performance

Step 6: Internal validation

Step 7: Model presentation

Subsection 2

Seven Steps for Model Development

Step 1: Problem definition and data inspection

What factors predict the endpoint?

What is the risk of the endpoint given a combination of factors?

Endpoint examples:

1. Functional rating

Step 1: Problem definition and data inspection

What factors predict the endpoint?

What is the risk of the endpoint given a combination of factors?

Endpoint examples:

1. Functional rating
2. Binary outcome like 30-day mortality, development of a disease

Step 1: Problem definition and data inspection

What factors predict the endpoint?

What is the risk of the endpoint given a combination of factors?

Endpoint examples:

1. Functional rating
2. Binary outcome like 30-day mortality, development of a disease
3. Time to an event, like death or disease progression

Step 1: Problem definition and data inspection

What is already known about the predictors?

Need close collaborations between the clinicians and the statisticians.

Step 1: Problem definition and data inspection

How were patients selected?

Step 1: Problem definition and data inspection

How to deal with treatment effects?

Include only the placebo group

Step 1: Problem definition and data inspection

Were the predictors reliably and completely measured?

Missing data

Step 1: Problem definition and data inspection

Is the endpoint of interest?

Preference for hard endpoints.

Step 2: Coding of predictors

Categorical variables: collapse categories with infrequent occurrence

Continous variables: in inital development, do not categorize (inefficient).

When you decide to roll-out, consider categorization for purposes of being user friendly *if* there is not too much of a loss of information.

Step 3: Model specification

How to choose? Stepwise procedures can perform poorly in small datasets.

Additivity association? Choice of interaction terms.

Step 3: Model specification

Tradeoff between

Simple robust model Finely-tuned model with many independent variables

Step 4: Model estimation

How to estimate those regression coefficients?

When using generalized linear model, use maximum likelihood for estimation

Step 4: Model estimation

Model may be overfitted, and need to "shrink" the regression coefficients towards zero, as they are typically too optimistic.

There are variety of ways to do this. When you are reading the literature, if someone says they used the

Penalized Maximum Likelihood method, or the Least Absolute Shrinkage and Selection Operator (LASSO)

just recognize that is what they are doing.

Step 5: Model Performance

Assess with some of the measures mentioned previously

- ROC curve

Step 5: Model Performance

Assess with some of the measures mentioned previously

- ROC curve
- Confusion matrix

Step 5: Model Performance

Assess with some of the measures mentioned previously

- ROC curve
- Confusion matrix
- Calibration

Step 5: Model Performance

Assess with some of the measures mentioned previously

- ROC curve
- Confusion matrix
- Calibration
- Discrimination ability

Step 5: Model Performance

Assess with some of the measures mentioned previously

- ROC curve
- Confusion matrix
- Calibration
- Discrimination ability
- AIC

Step 6: Model Validity

Internal validity: Validity of claims for the underlying population from which the data originated (reproducibility)

Use cross validation, bootstrap resampling to determine stability of - prediction variables - prediction quality

External validity: Generalizability of claims to 'plausibly related' populations

Looking at predictive ability for

- future patients (temporal validity)

Step 6: Model Validity

Internal validity: Validity of claims for the underlying population from which the data originated (reproducibility)

Use cross validation, bootstrap resampling to determine stability of - prediction variables - prediction quality

External validity: Generalizability of claims to 'plausibly related' populations

Looking at predictive ability for

- future patients (temporal validity)
- patients from another location (geographical validity)

Step 6: Model Validity

Internal validity: Validity of claims for the underlying population from which the data originated (reproducibility)

Use cross validation, bootstrap resampling to determine stability of - prediction variables - prediction quality

External validity: Generalizability of claims to 'plausibly related' populations

Looking at predictive ability for

- future patients (temporal validity)
- patients from another location (geographical validity)
- patients treated in another setting (strong external validation)

Step 7: Model presentation

Possibilities include:

- Regression formula

Step 7: Model presentation

Possibilities include:

- Regression formula
- Score charts

Step 7: Model presentation

Possibilities include:

- Regression formula
- Score charts
- Nomograms

Step 7: Model presentation

Possibilities include:

- Regression formula
- Score charts
- Nomograms
- Apps

Subsection 3

For better clinical prediction models

ABCD's of model validation

Four key measures in assessing validity of prediction models, related to calibration, discrimination, and clinical usefulness.

- Alpha

ABCD's of model validation

Four key measures in assessing validity of prediction models, related to calibration, discrimination, and clinical usefulness.

- Alpha
- Beta

ABCD's of model validation

Four key measures in assessing validity of prediction models, related to calibration, discrimination, and clinical usefulness.

- Alpha
- Beta
- Concordance statistic

ABCD's of model validation

Four key measures in assessing validity of prediction models, related to calibration, discrimination, and clinical usefulness.

- Alpha
- Beta
- Concordance statistic
- Decision Curve Analysis

Subsection 4

ABCD's of Model Validation

Alpha: Calibration in the Large

Calibration: agreement between observed and predicted, which is assessed graphically by plotting predicted (x-axis) v observed (y-axis)

Alpha: Compare mean of all predicted to mean of all observed

Beta: Calibration Slope

The calibration slope β is often estimated as smaller than 1 in a (relatively) small dataset.

Hence the need for cross-validation, bootstrapping estimates, and external validation

Concordance statistic: Discrimination

Discrimination: ability of the model to distinguish between a patient with an endpoint and one without the endpoint.

Quantification of this ability is via the concordance statistic, c.

The ROC plots sensitivity (true-positive rate) v 1-specificity (true-negative rate) for consecutive cutoffs for predicted risk. The concordance statistic, c, is measured by the AUC.

Decision curve analysis

Decision curve balances those with high risk to those with low risk.

Once a cutpoint has been chosen, summary measures include

- Sum of sensitivity and 1-specificity (naive)

Decision curve analysis

Decision curve balances those with high risk to those with low risk.

Once a cutpoint has been chosen, summary measures include

- Sum of sensitivity and 1-specificity (naive)
- Netbenefit: weighted sum of true- minus false-positives.

Section 4

Prediction

Subsection 1

Summary

Moral of the Story

Models are “easy” to develop (“just” involve computer)

Validation (especially external validation) is much more difficult, but must be done.