Introduction to Data Science - 1MS041

Benny Avelin

Department of Mathematics

HT 2022

Recall from last time

- The Data is our random variables $X = (X_1, \dots, X_n)$. The data is an observation of our Data.

Recall from last time

- The Data is our random variables $X = (X_1, \ldots, X_n)$. The data is an observation of our Data.
- A statistic is a function $\hat{\theta}$ from the data-space. We apply it on X, namely we are interested in $\hat{\theta}(X)$.

Recall from last time

- The Data is our random variables $X = (X_1, \ldots, X_n)$. The data is an observation of our Data.
- A statistic is a function $\hat{\theta}$ from the data-space. We apply it on X, namely we are interested in $\hat{\theta}(X)$.
- An estimator is a statistic that is supposed to "estimate" an unknown quantity, say θ^*. Therefore we can speak about bias

$$\text{bias} = \mathbb{E}[\hat{\theta}(X)] - \theta^*$$

Recall from last time

- The Data is our random variables $X = (X_1, \ldots, X_n)$. The data is an observation of our Data.
- A statistic is a function $\hat{\theta}$ from the data-space. We apply it on X, namely we are interested in $\hat{\theta}(X)$.
- An estimator is a statistic that is supposed to "estimate" an unknown quantity, say θ^*. Therefore we can speak about bias

$$\text{bias} = \mathbb{E}[\hat{\theta}(X)] - \theta^*$$

- A simple measure of performance is the standard deviation of the estimator, called, the standard error.

Recall from last time

- The Data is our random variables $X = (X_1, \ldots, X_n)$. The data is an observation of our Data.
- A statistic is a function $\hat{\theta}$ from the data-space. We apply it on X, namely we are interested in $\hat{\theta}(X)$.
- An estimator is a statistic that is supposed to "estimate" an unknown quantity, say θ^*. Therefore we can speak about bias

$$\text{bias} = \mathbb{E}[\hat{\theta}(X)] - \theta^*$$

- A simple measure of performance is the standard deviation of the estimator, called, the standard error.
- The risk of the estimator w.r.t the quadratic loss can be decomposed as

$$\mathbb{E}[(\hat{\theta}(X) - \theta^*)^2] = (\text{bias}(\hat{\theta}))^2 + (\text{se}(\hat{\theta}))^2$$

Recall from last time

- We also defined different modes of convergence
  - Almost sure convergence
  - Convergence in probability
  - Convergence in distribution (we did not define this)

Recall from last time

- We also defined different modes of convergence
  - Almost sure convergence
  - Convergence in probability
  - Convergence in distribution (we did not define this)
- an estimator is asymptotically consistent if it converges in probability to the true value.

Standard errors

Example

With the help of a balance scale we have measured the weights of two items with true weights m_1, m_2. We have measured $m_1, m_2, m_1 - m_2, m_1 + m_2$. Whenever we measure we make a measurement error with standard deviation σ. X_1, X_2, X_3, X_4 are the four measurements

Standard errors

Example

With the help of a balance scale we have measured the weights of two items with true weights m_1, m_2. We have measured $m_1, m_2, m_1 - m_2, m_1 + m_2$. Whenever we measure we make a measurement error with standard deviation σ. X_1, X_2, X_3, X_4 are the four measurements

- We can use X_1 as an estimate for m_1

Standard errors

Example

With the help of a balance scale we have measured the weights of two items with true weights m_1, m_2. We have measured $m_1, m_2, m_1 - m_2, m_1 + m_2$. Whenever we measure we make a measurement error with standard deviation σ. X_1, X_2, X_3, X_4 are the four measurements

- We can use X_1 as an estimate for m_1
- We could also use $(X_1 - X_2 + X_4)/2$ as an estimate for m_1

Standard errors

Example

With the help of a balance scale we have measured the weights of two items with true weights m_1, m_2. We have measured $m_1, m_2, m_1 - m_2, m_1 + m_2$. Whenever we measure we make a measurement error with standard deviation σ. X_1, X_2, X_3, X_4 are the four measurements

- We can use X_1 as an estimate for m_1
- We could also use $(X_1 - X_2 + X_4)/2$ as an estimate for m_1
- To estimate $m_1 + m_2$ we could use $(X_3 + X_4)/2$.

Standard errors

Example

With the help of a balance scale we have measured the weights of two items with true weights m_1, m_2. We have measured $m_1, m_2, m_1 - m_2, m_1 + m_2$. Whenever we measure we make a measurement error with standard deviation σ. X_1, X_2, X_3, X_4 are the four measurements

- We can use X_1 as an estimate for m_1
- We could also use $(X_1 - X_2 + X_4)/2$ as an estimate for m_1
- To estimate $m_1 + m_2$ we could use $(X_3 + X_4)/2$.
- To estimate $m_1 + m_2$ we could also use $(X_1 + X_2 + X_3)/2$

Standard errors

Example

A reliability system consists of two parallell circuits which break independently of each other and has the probabilities p_1 and p_2 to break during a week. The weekwise probability that the system breaks is thus $p_1 p_2$. We now test n such systems and find that the first circuit breaks x_1 times and the second system breaks x_2 times and that the full system has broken down x times. Two estimators for $p_1 p_2$ has been proposed.

$$p^* := \frac{x}{n}$$
$$\hat{p} := \frac{x_1}{n}\frac{x_2}{n}$$

Show that they are unbiased and compute the corresponding variances.

Maximum likelihood as risk minimization

Likelihood as a Risk minimization problem

Lets say we have a parametric model $\mathcal{E} = \{p_\alpha(z), \alpha \in \mathbb{R}^n\}$ for some family of densities p_α. Think of this parametrization (as an example)

$$p_\alpha(x) = \frac{1}{\alpha_2\sqrt{2\pi}} e^{-\frac{|z-\alpha_1|^2}{\alpha_2^2}}.$$

Likelihood as a Risk minimization problem

Lets say we have a parametric model $\mathcal{E} = \{p_\alpha(z), \alpha \in \mathbb{R}^n\}$ for some family of densities p_α. Think of this parametrization (as an example)

$$p_\alpha(x) = \frac{1}{\alpha_2\sqrt{2\pi}} e^{-\frac{|z-\alpha_1|^2}{\alpha_2^2}}.$$

Take the loss $L(z, \alpha) = -\ln p_\alpha(z)$ then the risk becomes

$$R(\alpha) = -\int \ln(p_\alpha(z)) p_{\alpha^*}(z) dx$$

Likelihood as a Risk minimization problem

Lets say we have a parametric model $\mathcal{E} = \{p_\alpha(z), \alpha \in \mathbb{R}^n\}$ for some family of densities p_α. Think of this parametrization (as an example)

$$p_\alpha(x) = \frac{1}{\alpha_2\sqrt{2\pi}} e^{-\frac{|z-\alpha_1|^2}{\alpha_2^2}}.$$

Take the loss $L(z,\alpha) = -\ln p_\alpha(z)$ then the risk becomes

$$R(\alpha) = -\int \ln(p_\alpha(z)) p_{\alpha^*}(z) dx$$

If we let Z be a random variable with law p_{α^*} then we can write the above as

$$R(\alpha) = \mathbb{E}[-\ln(p_\alpha(Z))]$$

Estimating the risk

An estimator of the risk is the so called empirical risk. Given a sequence of i.i.d. random variables $Z_1, \ldots, Z_n$ sampled from p_{α^*} the empirical Risk is

$$\hat{R}(\alpha) = -\frac{1}{n}\sum_{i=1}^{n} \ln(p_\alpha(Z_i)).$$

Estimating the risk

An estimator of the risk is the so called empirical risk. Given a sequence of i.i.d. random variables $Z_1, \ldots, Z_n$ sampled from p_{α^*} the empirical Risk is

$$\hat{R}(\alpha) = -\frac{1}{n}\sum_{i=1}^{n} \ln(p_\alpha(Z_i)).$$

The quantity

$$-n\hat{R}(\alpha) = \sum_{i=1}^{n} \ln(p_\alpha(Z_i))$$

is the well known **log-likelihood**.

Lets consider a simple case

Now consider the parametrization $N(0, \alpha^2)$, i.e.

$$p_\alpha(x) = \frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{|z|^2}{2\alpha^2}}.$$

Lets consider a simple case

Now consider the parametrization $N(0, \alpha^2)$, i.e.

$$p_\alpha(x) = \frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{|z|^2}{2\alpha^2}}.$$

Lets write the empirical risk

$$\begin{aligned}\hat{R}(\alpha) &= -\frac{1}{n}\sum_{i=1}^{n} \ln(p_\alpha(Z_i)) = -\frac{1}{n}\sum_{i=1}^{n} \ln(\frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{|Z_i|^2}{2\alpha^2}}.) \\ &= \ln(\alpha) + \frac{1}{n}\sum_{i=1}^{n} \frac{|Z_i|^2}{2\alpha^2} + c\end{aligned}$$

Lets find the critical point

$$\frac{d}{d\alpha}\hat{R}(\alpha) = 0$$

Lets find the critical point

$$\frac{d}{d\alpha}\hat{R}(\alpha) = 0$$

Computing the derivative we get

$$\frac{d}{d\alpha}\hat{R}(\alpha) = \frac{1}{\alpha} - \frac{1}{n}\sum_{i=1}^{n}\frac{|Z_i|^2}{\alpha^3} = 0$$

Multiplying by α^3 on both sides and moving over gives

$$\alpha^2 = \frac{1}{n}\sum_{i=1}^{n}|Z_i|^2.$$

Thus, the empirical variance has the minimal log-risk.

Likelihood and regression / logistic regression

Lets consider some certain cases

Consider now a density $f_{X,Y}(x, y)$ for the pair (X, Y), then we can compute

$$\ln(f_{X,Y}(x, y)) = \ln(f_{Y|X}(y \mid x) f_X(x)) = \ln(f_{Y|X}(y \mid x)) + \ln(f_X(x))$$

Lets consider some certain cases

Consider now a density $f_{X,Y}(x,y)$ for the pair (X,Y), then we can compute

$$\ln(f_{X,Y}(x,y)) = \ln(f_{Y|X}(y \mid x) f_X(x)) = \ln(f_{Y|X}(y \mid x)) + \ln(f_X(x))$$

We now make the assumption that f_X is a fixed density, that we will not care about and instead assume that $f_{Y|X}$ is part of a parametrized family, i.e. $f_{Y|X} = p_{\alpha^*,X}$ for some α^*.

Lets consider some certain cases

Consider now a density $f_{X,Y}(x,y)$ for the pair (X,Y), then we can compute

$$\ln(f_{X,Y}(x,y)) = \ln(f_{Y|X}(y \mid x) f_X(x)) = \ln(f_{Y|X}(y \mid x)) + \ln(f_X(x))$$

We now make the assumption that f_X is a fixed density, that we will not care about and instead assume that $f_{Y|X}$ is part of a parametrized family, i.e. $f_{Y|X} = p_{\alpha^*,X}$ for some α^*.

- $p_{\alpha^*,X} = N(\alpha_1 X + \alpha_2, \alpha_3^2)$, Linear regression
- $p_{\alpha^*,X} = \text{Bernoulli}(G(\alpha_1 X + \alpha_2))$,

  $$G(x) = \frac{1}{1+e^{-x}}$$

  Logistic regression

Lets derive the case of linear regression

The empirical risk is thus

$$-\frac{1}{n}\sum_{i=1}^{n}\ln(p_{\alpha^*,X_i}(Y_i)) = -\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{1}{\alpha_3^2\sqrt{2\pi}}e^{-\frac{1}{2\alpha_3^2}(Y_i-(\alpha_1 X_i+\alpha_2))^2}\right)$$

Lets derive the case of linear regression

The empirical risk is thus

$$-\frac{1}{n}\sum_{i=1}^{n}\ln(p_{\alpha^*,X_i}(Y_i)) = -\frac{1}{n}\sum_{i=1}^{n}\ln(\frac{1}{\alpha_3^2\sqrt{2\pi}}e^{-\frac{1}{2\alpha_3^2}(Y_i-(\alpha_1 X_i+\alpha_2))^2})$$

Thus

$$-\frac{1}{n}\sum_{i=1}^{n}\ln(p_{\alpha^*,X_i}(Y_i)) = \ln(\alpha_3^2) + \frac{1}{2\alpha_3^2}\frac{1}{n}\sum_{i=1}^{n}(Y_i-(\alpha_1 X_i+\alpha_2))^2.$$

Lets derive the case of linear regression

The empirical risk is thus

$$-\frac{1}{n}\sum_{i=1}^{n}\ln(p_{\alpha^*,X_i}(Y_i)) = -\frac{1}{n}\sum_{i=1}^{n}\ln(\frac{1}{\alpha_3^2\sqrt{2\pi}}e^{-\frac{1}{2\alpha_3^2}(Y_i-(\alpha_1X_i+\alpha_2))^2})$$

Thus

$$-\frac{1}{n}\sum_{i=1}^{n}\ln(p_{\alpha^*,X_i}(Y_i)) = \ln(\alpha_3^2) + \frac{1}{2\alpha_3^2}\frac{1}{n}\sum_{i=1}^{n}(Y_i-(\alpha_1X_i+\alpha_2))^2.$$

If we ignore α_3 then the best α_1, α_2 can be found by solving

$$(\alpha_1^*, \alpha_2^*) = \arg\min_{\alpha_1,\alpha_2}\frac{1}{n}\sum_{i=1}^{n}(Y_i-(\alpha_1X_i+\alpha_2))^2$$

Conclusion

Conclusion

Thus linear regression gives rise to the quadratic loss!

Logistic regression

We can do similar reasoning for the case when we assume that the conditional distribution is $p_{\alpha^*,X} = \text{Bernoulli}(G(\beta_0 + \beta_1 X))$, where

$$G(x) = \frac{1}{1 + e^{-x}}, \quad \text{logistic function.}$$

Logistic regression

If we call $p(X) = G(\beta_0 + \beta_1 X)$ then

$$
\begin{aligned}
-\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)) &= -\sum_{i=1}^{n} \ln(p(X_i)^{Y_i}(1 - p(X_i))^{1-Y_i}) \\
&= -\sum_{i=1}^{n} (Y_i \ln(p(X_i)) + (1 - Y_i) \ln(1 - p(X_i)))
\end{aligned}
$$

Logistic regression

If we call $p(X) = G(\beta_0 + \beta_1 X)$ then

$$
\begin{aligned}
-\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)) &= -\sum_{i=1}^{n} \ln(p(X_i)^{Y_i}(1 - p(X_i))^{1-Y_i}) \\
&= -\sum_{i=1}^{n} (Y_i \ln(p(X_i)) + (1 - Y_i) \ln(1 - p(X_i)))
\end{aligned}
$$

Thus all we have to do is to minimize

$$
-\sum_{i=1}^{n} (Y_i \ln(p(X_i)) + (1 - Y_i) \ln(1 - p(X_i)))
$$

Logistic regression

If we call $p(X) = G(\beta_0 + \beta_1 X)$ then

$$
\begin{aligned}
-\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)) &= -\sum_{i=1}^{n} \ln(p(X_i)^{Y_i}(1 - p(X_i))^{1-Y_i}) \\
&= -\sum_{i=1}^{n} (Y_i \ln(p(X_i)) + (1 - Y_i)\ln(1 - p(X_i)))
\end{aligned}
$$

Thus all we have to do is to minimize

$$
-\sum_{i=1}^{n} (Y_i \ln(p(X_i)) + (1 - Y_i)\ln(1 - p(X_i)))
$$

This is a loss function

$$
L(a, b) = b\ln(a) + (1 - b)\ln(1 - a)
$$

called the binary cross entropy, or simply the log-loss.

Logistic regression: numerical aspects

We can often simplify this loss, like for our function G. Note that

$$\ln(p(X_i)) = \ln(1/(1+e^{-(\beta_0+\beta_1 X_i)})) = -\ln(1+e^{-(\beta_0+\beta_1 X_i)})$$
$$\ln(1-p(X_i)) = \ln(1-1/(1+e^{-(\beta_0+\beta_1 X_i)})) = -\ln(1+e^{\beta_0+\beta_1 X_i}).$$

Thus the only thing that changes is the sign of the exponent, so if we write $Z_i = 2Y_i - 1$ then $Z_i = 1$ if $Y_i = 1$ and $Z_i = -1$ if $Y_i = 0$ and we can write

$$-\sum_{i=1}^{n} \ln\left(p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}\right) = \sum_{i=1}^{n} \ln(1+e^{-Z_i(\beta_0+\beta_1 X_i)}).$$