Statistical Machine Learning

Linus Falk

November 9, 2022

Contents

Lecture 2: Linear regreesion

wednesday 02 nov 10:15

1 Linear regression discussion points

Different reasons we look at the square error, will be discussed.
Cost landscape in parameter space.
minimizes can be hard to find

- In what type of problems is the squared prediction error an unsuitable performance measure:

It symmetric, we maybe don't want to under overestimate the outcome/prediction. Think of medication or battery. Another way yo do it pinball loss. two alternatives

$$L(x, y, \theta) = \left[Pinballloss\right] (y - f(x, \theta))\alpha, y \geq f(x, \theta) or \\ (y - f(x, \theta))(1 - \alpha), y < f(x, \theta)) \tag{1}$$

- How would you visualize a regression model with two inputs.

A plane for a linear regression model. A surface lives in a subspace of the d parameters space.

- **When is there a unique linear regression model that minimizes the average squared-error loss**

IF we get to little data we cant generate a unique model for an example. Think of just having one data point and fitting a line to it. All interpolate the training data
Necessary for uniqueness that $n \geq (d + 1)$, (not sufficient)
Linear regression model

$$\overline{X} = \begin{bmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{bmatrix} \tag{2}$$

$X^T X$ is invertible. (d+1) when they are linearly independent we can find a unique solution. (rand($\overline{X}$) = d+1)

- 4

Parameter space Numerical search e.g. gridding. Evaluate J at every gridpoint. Might not find the actual minimum. Might work when d i small ≤ 3. Another way is gradient search.

- Consider alternative ways to regularize the least-square method.

$J(\theta) = 1/nnorm(\overline{y} - \overline{X\theta})$
Positive quadratic function in θ (convex) local minimum $\rightarrow$ global minimum of J(θ)
Local min (*) $\nabla_0 J(\theta) = 0 <-> \overline{X^T X} = X^T \overline{y}$
**Notes if small amount of data, choose fewer d, but which one to choose?
Different norms for example norm 2 for ridge regression affects the cost. Lasso is another example using norm 1.
Regularization reduces the sensitivity.

- How does one interpret the 'best' regression function?

Finding the balancing point of the distribution for each point.

- In what types of problem can it still produce poor performance

Balancing point between two "hills" give a prediction between them. Data will never come there. Multi modal distribution. The population splits into two parts. In the blood pressure it could be difference between male and female there could therefore be necessary add a new feature.

Family of Gaussian distributions

Why not model the best regression function

Lecture 3: Classification

tuesday 08 nov 10:15

2 Classification

We return to the example with blood pressure and cholesterol. We got the patient data

- x_1 change in blood pressure during exercise
- x_2 total cholesterol
- y stroke within 5 years {-1, +1}

Given this data x_*, we want to predict strokes within five years y_* using the training data set.

The expected new error of a model

At an unknown future point, the miss classification error of the model is

$$\mathbf{1}\{y_* \neq f(x_*;\theta)\} \tag{3}$$

This is easy to interpret and easy to analyze. We now want to find θ to minimize our expected error:

$$\mathbf{E}\left[\mathbf{1}\{y_* \neq f(x_*;\theta\}\right] \tag{4}$$

but as in previous case p(x,y) is Unknown! We return to:

Learning a linear classifier

We can span a linear half plane by

$$\{x \quad x^T\theta \quad \geq 0\} \tag{5}$$

for a given vector θ

Our linear classifier:

$$f(x;\theta) = \text{sign}(x^T\theta) = \begin{cases} +1, & x^T\theta \geq 0 \\ -1, & x^T\theta < 0 \end{cases} \tag{6}$$

Learn a model, the θ parameters by minimizing the cos function (average loss)

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i; \theta) \tag{7}$$

with the aim to reduce the expected error from a new observation (new error)

$$\mathbf{E}_{new\theta}) = \mathbf{E}_* \left[E(f(x_*;\theta), y_*)\right] \tag{8}$$

Computational challenge

This is a challenge to compute and it doesn't take the consideration of how close this classifier lines is put to observations. Average loss:

$$J(\theta) = \frac{1}{n}\sum_{i=0}^{n} L(x_i, y_i; \theta), \text{where} L(x_i, y_i; \theta) = \mathbf{1}\{y \neq \text{sign}(x^T\theta)\} \tag{9}$$

Loss function and classifier margin

The margin of the (linear) classifier is defined as:

$$y \times x^T\theta \tag{10}$$

Comparing the miss classification loss functions. The second, logistic loss:

$$\begin{gathered} L(x, y; \theta) = \begin{cases} 0, & yx^T \geq 0 \\ 1, & yx^T\theta < 0 \end{cases} \\ L(x, y; \theta) = \ln\left[1 + e^{-yx^T\theta}\right] \end{gathered} \tag{11}$$

The linear classifier f(x; θ) = sign($x^t\theta$) learned by minimizing

$$J(\theta)\frac{1}{n}\sum_{i=0}^{n} L(x_i, y_i; \theta) \tag{12}$$

where the logistic loss is a convex function of θ(can be minimized). Like in previous regression problems we can use regularization to reduce the sensitivity of the learned model $\hat{\theta}$. We may regularize the cost by:

$$J(\theta) + \lambda||\theta||_2^2 \tag{13}$$

Discussion points* Classification

- **n what type of problems is the missclassification error an unsuitable performance measure?**
  Bad error method when it's not symmetric, could be better to classify risk of stroke instead of not. In medicine there is often a big asymmetry. Two types of error.

- **How does the missclassification loss of a linear classifier change if you rescale the parameters θ ?**
  It doesn't change, linear.

- **How does the logistic loss increase with the (negative) margin of missclassified points.**

$$\begin{aligned} L(x,y;\theta) &= \ln\left[1+e^{-yx^T\theta}\right] \\ &\approx -y \times x^T\theta \end{aligned} \tag{14}$$

  Linearly for very "negative" numbers. LOOK at this again

- **The logistic loss is convex in θ. What does this imply in parameter space?**
  Two theories, two universes, convex and not convex in optimization.

$L(x,y;\theta)$ convex function $\rightarrow J(\theta)\frac{1}{n}\sum_i^n L(x,y;\theta)$
also convex func in $\theta J(w\theta+(1-w)\theta') \leq wJ(\theta)+(1-w)J(\theta')$
$\rightarrow$ all minima of J, arg min J. Form a convex set

The best classifier

Search through all functions f(x) that can minimize the expected value of an new observation:

$$E_{new} = \mathbf{E}_*\left[\mathbf{1}\{y_* \neq f(x_*)\}\right] \tag{15}$$

The conditional distribution for y=+1: p(y=1|x)

The model that minimizes $E_{new} = \mathbf{E}_*\left[\mathbf{1}\{y_* \neq f(x_*)\}\right]$ is given by the conditional distribution:

$$\mathrm{f}_0(x) = \mathrm{argmax} p(y|x) \tag{16}$$

Discussion point* The best classifier

- **The plot illustrates p(y = 1|x), how does p(y = -1|x) look? How does p(x) look?**

$$p(y=-1|x) = \left[\text{complementary event}\right] = 1 - p(y=1|x) \tag{17}$$

  p(x) probability were there might be data points.

- **How does one interpret the 'best' classifier?**
  what does this arg max p(y|x) mean. y $\in \{-1, 1\}$

$$\begin{cases} +1, & p(1|x) > p(-1|x) \\ -1, & p(1|x) \leq p(-1|x) \end{cases} \tag{18}$$

Alternative 1:

$$p(1|x) > 1 - p(1|x) \leftrightarrow p(1|x) > \frac{1}{2} \tag{19}$$

Alternative 2, threshold:

$$\frac{p(1|x)}{(0|x)} > 1 \rightarrow \text{threshold} \ \frac{p(1|x)}{(0|x)} > \text{T} \tag{20}$$

- **In what types of problems can it still produce poor performance?** Same answer as before, medical example. Missing out stroke patients is serious.

Alternative loss: the likelihood perspective

The best classifier $\mathrm{f}_0(\mathrm{x})$ depends on: p(y|x). Let us now model that directly. We have the family of distribution models:

$$p(y|x;\theta) = \frac{e^{yx^T\theta}}{1+e^{yx^T\theta}} \tag{21}$$

Which gives a model for $\mathrm{f}_0(\mathrm{x})$ also known as logistic regression

How surprising is the training data?

For the model θ, the surprise of the training data point $(x_i, y_i$ is given by:

$$L(x_i, y_i; \theta) = -\ln p(y_i|x_i;\theta) \tag{22}$$

which is also known as the negative log-likelihood loss

$$\hat{\theta} = \arg\min \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i; \theta) \tag{23}$$

that is the maximum likelihood model of conditional distribution

$$p(y|x;\hat{\theta}) = \left[1 + e^{-yx^T\hat{\theta}}\right]^{-1} \tag{24}$$

For logistic distribution model, $\hat{\theta}$(training data)) matches the logistic loss minimizer.

Discussion points* Alternative loss

- **How does one interpret the parameters of a linear model?** $f(x;\theta) = x^T\theta \rightarrow \theta_0 + \theta_1 x_1 \dots$

  what mean θ_1 models association with between x_1 and y when all other x_i are fixed. Logistic distrubution:

  $$ln(\frac{p(y=1|x)}{p(y=-1|x)}) = x^T\theta \tag{25}$$

  θ affects the low odds for 'stroke'.

Special case, classes are linearly separable.

$$L = \ln \left[1 + e^{-yx^T\theta}\right] \tag{26}$$

Can make the loss smaller and smaller by blowing up the parameters. Making it sharper and sharper. J(θ) has no minimal point.

- **Why is the surprisal equivalent to the logistic loss?**

$$p(y|x) = \frac{e^{yx^T\theta}}{1 + e^{yx^T\theta}} = \frac{1}{e^{-yx^T\theta} + 1} = \left[1 + e^{-yx^T\theta}\right]^{-1} \tag{27}$$

$$L = -\ln p(x, y; \theta) = -\ln \left[1 + e^{-yx^T\theta}\right]^{-1} = \ln \left[\ldots\right] \tag{28}$$

negative log likelihood

- **What are some advantages/disadvantages of using parametric distributional modelling?**
Instead of hard predictions we can get a probability. Some times called soft classification. The cons are : its an uncalibrated model. Gap in predictions.

- **How does one extend the model to handle M > 2 classes?**
The idea how to do it:

$$\begin{aligned} y &\in 1, 2, \ldots, M \\ *\ln\frac{p(y = 1|x)}{p(y = M|x)} &= x^T\theta_1 \\ \ln\frac{p(y = 2|x)}{p(y = M|x)} &= x^T\theta_2 \\ &\vdots \\ \ln\frac{p(y = M-1|x)}{p(y = M|x)} &= x^T\theta_{M-1} \end{aligned} \tag{29}$$

$$* * 1 \sum_{y=1}^{M} p(y|x)) = 1 \tag{30}$$

$$(*) \to p(y = m|x) = p(y = M|x)e^{x^T\theta_m} = \frac{e^{x^T\theta_m}}{1 + \sum_{k=1}^{M-1} e^{x^T\theta_k}} \tag{31}$$

Lecture 4: Classification

wednesday 09 nov 10:15

3 Multivariate Gaussian density

The p-dimensional Gaussian probability density function with mean vector μ and covariance matrix Σ is,

$$N(x|\mu, \Sigma = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{32}$$

where $\mu : p \times 1$ vector and $\Sigma : p \times p$ positive definite matrix.
If we let $\mathbf{x} = (x_1, \dots, x_p)^T tildeN(\mu, \Sigma)$

- μ_j is the mean of x_j
- Σ_{jj} is the variance of x_j
- $\Sigma_{ij}(i \neq j)$ is the covariance between x_i and x_j

4 Linear discriminant analysis, LDA

Here we need,

- The prior class probabilities π_m, p(y=m), m $\in \{1, ..., m\}$
- The conditional probability densities of the input x, $f_m(x)$, p(x |y=m) for each class m.

This will give us the model:

$$g_m(x) = \frac{\pi_m f_m(x)}{\sum_{m=1}^{M} \pi_m f_m(x)} \tag{33}$$

For **first task** a natural estimator is the proportion of training samples in the m th class.:

$$\hat{pi}_m = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}\{y_m = m\} = \frac{n_m}{n} \tag{34}$$

where n is the size of the training set and n_m the number of training samples of class m.

for the **second task** a simple model is to assume that $f_m(x)$ is a multivariate normal density with mean vector μ_m and covariance matrix Σ_m

$$f_m(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_m)^T \sum_m^{-1}(x-\mu_m)} \tag{35}$$

if we further assume that all classes share the same covariance matrix,

$$\Sigma = \text{def}\Sigma_1 = \ldots = \Sigma_M \tag{36}$$

the remaining parameters of the model are: $\mu_1, \mu_2, \ldots, \mu_M, \Sigma$

These parameters are naturally estimated as the sample means and sample covariance, respectively:

$$\begin{gathered}
\hat{\mu}_m = \frac{1}{n_m} \sum_{i:y_i=m} x_i \quad m = 1, \ldots, M \\
\hat{\Sigma} = \frac{1}{n-M} \sum_{m=1}^{M} \sum_{i:y_i=m} (x_i - \hat{\mu}_m)(x_i - \hat{\mu}_m)^T
\end{gathered} \tag{37}$$

Modeling the class probabilities using the normal assumptions and these parameter estimates is referred to as **Linear Discriminant Analysis (LDA)**

The LDA classifier assigns a test input x to class m for which

$$\hat{\delta}_m = x^T \hat{\Sigma}^{-1} \hat{\mu}_m - \frac{1}{2} \hat{mu}_m + \log \hat{\pi}_m \tag{38}$$

is largest, where

$$\begin{aligned} \hat{\pi}_m &= \frac{n_m}{n}, \quad m = 1, \ldots, M \\ \hat{\mu}_m &= \frac{1}{n_m} \sum_{i:y_i=m} (x_i - \hat{\mu}_m)(x_i - \hat{mu}_m)^T \\ \hat{\Sigma} &= \frac{1}{n-M} \sum_{m=1}^{M} \sum_{i:y_i=m} (x_i - \hat{\mu}_m)(x_i - \hat{\mu}_m)^T \end{aligned} \tag{39}$$

Example: Difference between LDA and QDA

- **If the optimal boundary is linear, do we expect LDA or QDA to perform better on the training set? What do we expect on the test set?**
  We can always assume that QDA performs better to the test set since it is more flexible. If the optimal decision boundary is linear LDA will perform better on test data since it would not overfit.

- **If the optimal decision boundary is nonlinear, do we expect LDA or QDA to perform better on the training set? What do we expect on the test set?**
  If the optimal decision boundary is non linear We expect QDA to perform better on the test set.

- **In general, as the sample size n increases, do we expect the test error rate of QDA relative to LDA to increase, decrease or be unchanged? Why?**
  We can assume that the QDA error rate will reduce when n increases, with more n it will come close and closer to the optimal decision boundary while when n is small the risk of overfitting is increasing.

- **True or false: Even if the optimal decision boundary for a given problem is linear, we will probably achieve a smaller test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.**
  False if n is small we have a risk of overfitting

4.1 Quadratic discrimination analysis

Question: do we have to assume a common covariance matrix? **No**, estimating a separeate covariance matrix for each class leads to the method: Quadratic discrimination analysis or QDA for short. Which one to chose has to do with the bias-variance trade of or in other words the risk of over or under-fitting. Comparing LDA to QDA:

- has more parameters
- is more flexible
- has higher risk of overfitting (large variance)