Regression II

COMP9417, 23T2

Section 1

Stats, stats, stats . . .

Probability Distribution

A probability distribution represents the probability we see a value x in a sample X. We denote this as $P(X = x)$.

Definition

- Probability *mass* function applies to discrete X
- Probability *density* function applies to continuous X

Expected Values

An expected value (denoted $\mathbb{E}$) represents the weighted average of the probability distribution. This is typically seen as the value the random variable will converge to over time if sampled randomly.

For a discrete random variable, the form for an expected value is as follows:

$$\mathbb{E}(X) = \sum_{x \in X} xP(X = x)$$

In the continuous case, where $f(x)$ is the probability density function:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

General rules

For random variables X, Y and a constant c

- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[cX] = c\mathbb{E}[X]$

Example

Problem: Model the probability mass function and find the expected value of the roll of a dice.

Example

Problem: Model the probability mass function and find the expected value of the roll of a dice.

- $P(X = 1) = \frac{1}{6}$, $P(X = 2) = \frac{1}{6}$, $\ldots$, $P(X = 6) = \frac{1}{6}$

Example

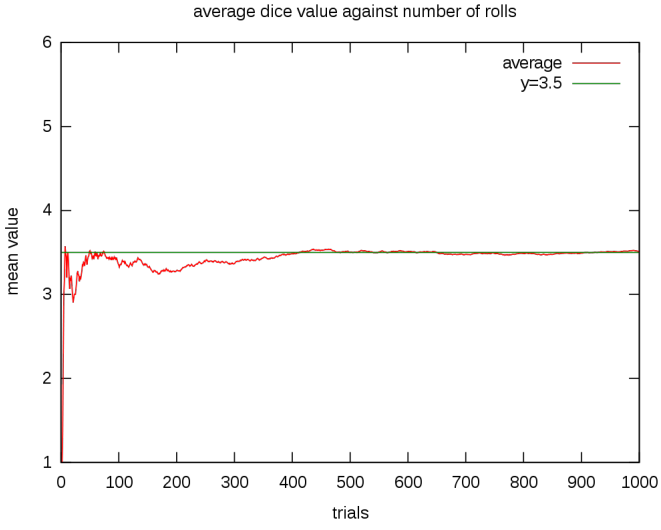
Problem: Model the probability mass function and find the expected value of the roll of a dice.

- $P(X = 1) = \frac{1}{6}$, $P(X = 2) = \frac{1}{6}$, $\ldots$, $P(X = 6) = \frac{1}{6}$

For the expected value:

$$\mathbb{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$$

This means that if we roll a dice, overtime the average dice value will converge to 3.5.


average dice value against number of rolls
average
y=3.5
mean value
trials
1
2
3
4
5
6
0
100
200
300
400
500
600
700
800
900
1000

We typically assume our samples are i.i.d (independent and identically distributed), helping us reduce the complexity of the problem and apply statistically supported conclusions.

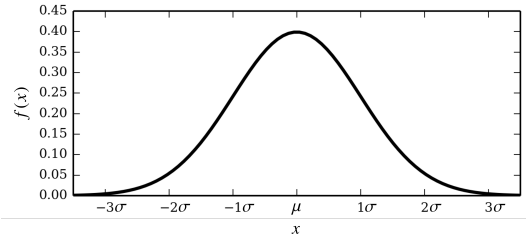
Gaussian Distribution

A standard probability distribution is the Gaussian, where:

$$\theta = (\mu, \sigma^2), \quad \mu \in \mathbb{R}, \sigma > 0$$

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

We typically write $X \sim \mathcal{N}(\mu, \sigma^2)$ to say X is normally distributed with mean μ and variance σ^2.



Maximum Likelihood Estimation

Maximum Likelihood Estimation

Maximum likelihood estimation is the process of estimating the parameters of a distribution of sample data by maximising the overall likelihood of the samples occuring in the distribution.

$$
\begin{aligned}
\text{Prob of observing } X_1, \ldots, X_n &= \text{Prob of observing } X_1 \times \cdots \times \text{Prob of observing } X_n \\
&= p_\theta(X_1) \times \cdots \times p_\theta(X_n) \\
&= \prod_{i=1}^{n} p_\theta(X_i) \\
&=: L(\theta) \qquad \text{this is our likelihood function}
\end{aligned}
$$

To make life easier, we typically work with the log of the likelihood function (log-likelihood). As $\log$ is a strictly increasing function, the maximisation of $L(\theta)$ and $\log(L(\theta))$ give us the same result.

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p_\theta(X_i) \\
\log(L(\theta)) &= \log \prod_{i=1}^{n} p_\theta(X_i) \\
&= \sum_{i=1}^{n} \log p_\theta(X_i)
\end{aligned}
$$

This makes differentiating, and therefore maximising much simpler.

Section 2

1 (a, b)

1a

Problem: Given $X_1, \cdots, X_n \sim N(\mu, 1)$, find $\hat{\mu}_{\mathsf{MLE}}$.

1a

Problem: Given $X_1, \cdots, X_n \sim N(\mu, 1)$, find $\hat{\mu}_{\text{MLE}}$.

First, we define our likelihood function:

$$\begin{aligned}
\log L(\mu) &= \log\left(\prod_{i=1}^{n} p_\theta(X_i)\right) \\
&= \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i - \mu)^2\right)\right) \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2
\end{aligned}$$

Next, we differentiate with respect to our parameter μ,

Next, we differentiate with respect to our parameter μ,

$$\begin{aligned}\frac{\partial \log L(\mu)}{\partial \mu} &= \sum_{i=1}^{n}(X_i - \mu) \\ &= \sum_{i=1}^{n} X_i - n\mu\end{aligned}$$

$\frac{\partial \log L(\mu)}{\partial \hat{\mu}} = 0$ at the maximum. So,

$$\sum_{i=1}^{n} X_i - n\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\mu} = \bar{X}$$

1b

Problem: Given $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, find $\hat{p}_{\text{MLE}}$.

The Bernoulli distribution models processes with 2 outcomes (eg. a coin toss).

$$P(X = k) = p^k(1-p)^{1-k}, \qquad k = 0, 1 \quad p \in [0, 1]$$

1b

Problem: Given $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, find $\hat{p}_{\text{MLE}}$.

The Bernoulli distribution models processes with 2 outcomes (eg. a coin toss).

$$P(X = k) = p^k(1-p)^{1-k}, \qquad k = 0, 1 \quad p \in [0, 1]$$

First, we define our likelihood function:

$$\begin{aligned}
\log L(p) &= \log\left(\prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}\right) \\
&= \sum_{i=1}^{n} \log p^{X_i} + \sum_{i=1}^{n} \log(1-p)^{1-X_i} \\
&= n\bar{X} \log p + n(1-\bar{X}) \log(1-p)
\end{aligned}$$

Next, we differentiate with respect to our parameter p,

Next, we differentiate with respect to our parameter p,

$$\frac{\partial \log L(p)}{\partial p} = \frac{n\bar{X}}{p} - \frac{n(1-\bar{X})}{1-p}$$

$\frac{\partial \log L(p)}{\partial \hat{p}} = 0$ at the maximum. So,

Next, we differentiate with respect to our parameter p,

$$\frac{\partial \log L(p)}{\partial p} = \frac{n\bar{X}}{p} - \frac{n(1-\bar{X})}{1-p}$$

$\frac{\partial \log L(p)}{\partial \hat{p}} = 0$ at the maximum. So,

$$\begin{aligned}
\frac{n\bar{X}}{\hat{p}} - \frac{n(1-\bar{X})}{1-\hat{p}} &= 0 \\
n\bar{X} - n\bar{X}\hat{p} &= n(1-\bar{X})\hat{p} \\
\hat{p}(n(1-\bar{X}) + n\bar{X}) &= n\bar{X} \\
\hat{p} &= \bar{X}
\end{aligned}$$

Section 3

Bias & Variance

Bias

The bias of an estimator represents its theoretical error. This theoretical error is the distance of the expected value of the predicted estimator away from the true parameter. We take the expected value for a representation of the estimate over an infinitely large dataset.

In the case of a model, this represents the error of the model on the training set.

There will be a more in-depth discussion of this later on in the course.

$$\text{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Variance

The variance of an estimator refers to how different or *variable* it is over different data. Again, we take the expected value to find the converging value over a theoretically infinitely large dataset.

When applying this to a model, we'll discover that a large variance and a low bias typically means that our model has *overfit* the training set.

$$\text{var}(\hat{\theta}) = \mathbb{E}(\theta - \mathbb{E}(\hat{\theta}))^2$$

Section 4

2 (a, b, c)

2a

Problem: Find the bias and variance of $\hat{\mu}_{\text{MLE}}$ for $X \sim N(\mu, 1)$.

We know that $\hat{\mu}_{\text{MLE}} = \bar{X}$. So,

2a

Problem: Find the bias and variance of $\hat{\mu}_{\text{MLE}}$ for $X \sim N(\mu, 1)$.

We know that $\hat{\mu}_{\text{MLE}} = \bar{X}$. So,

$$\begin{aligned}
\text{bias}(\bar{X}) &= \mathbb{E}(\bar{X}) - \mu \\
&= \frac{1}{n}\mathbb{E}\left(\sum_{i=0}^{n} X_i\right) - \mu \\
&= \frac{1}{n}\sum_{i=0}^{n}\mathbb{E}(X_i) - \mu \\
&= \frac{1}{n}n\mu - \mu \\
&= 0
\end{aligned}$$

Now, for the variance:

Now, for the variance:

$$\begin{aligned}\mathsf{var}(\hat{\mu}) &= \mathsf{var}(\bar{X}) \\ &= \mathsf{var}\left(\frac{1}{n}\sum_{i=0}^{n}\mathbb{E}(X_i)\right)\end{aligned}$$

Now, for the variance:

$$
\begin{aligned}
\mathsf{var}(\hat{\mu}) &= \mathsf{var}(\bar{X}) \\
&= \mathsf{var}\left(\frac{1}{n}\sum_{i=0}^{n}\mathbb{E}(X_i)\right) \\
&= \frac{1}{n^2}\sum_{i=0}^{n}\mathsf{var}(X_i) \\
&= \frac{1}{n^2}n = \frac{1}{n}
\end{aligned}
$$

2b

Find the bias and variance of $\hat{p}_{\text{MLE}}$ for $X \sim \text{Bernoulli}(p)$.

We know that $\hat{p}_{\text{MLE}} = \overline{X}$, so

$$\begin{aligned} \text{bias}(\hat{p}_{\text{MLE}}) &= \mathbb{E}(\hat{p}_{\text{MLE}}) - \mu \\ &= \overline{X} - \mu \\ &= 0 \end{aligned}$$

Now, for the variance:

Note that the variance of an r.v. with a Bernoulli(p) dist. is $p(1-p)$

Now, for the variance:

Note that the variance of an r.v. with a Bernoulli(p) dist. is $p(1-p)$

$$\begin{aligned}
\mathsf{var}(\hat{p}_{\mathsf{MLE}}) &= \mathsf{var}(\bar{X}) \\
&= \mathsf{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{var}\left(x_i\right) \\
&= \frac{1}{n^2} np(1-p) \\
&= \frac{p(1-p)}{n}
\end{aligned}$$

2c

Perform bias-variance decomposition i.e prove that MSE($\hat{\theta}$) =bias($\hat{\theta}$)^2 + var($\hat{\theta}$).

We are given the definition MSE($\hat{\theta}$) $= \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$.

2c

Perform bias-variance decomposition i.e prove that MSE$(\hat{\theta})$ =bias$(\hat{\theta})^2$ + var$(\hat{\theta})$.

We are given the definition MSE$(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$.

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

2c

Perform bias-variance decomposition i.e prove that MSE($\hat{\theta}$) =bias($\hat{\theta}$)^2 + var($\hat{\theta}$).

We are given the definition MSE($\hat{\theta}$) $= \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$.

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

2c

Perform bias-variance decomposition i.e prove that MSE$(\hat{\theta})$ =bias$(\hat{\theta})^2$ + var$(\hat{\theta})$.

We are given the definition MSE$(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$.

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

Finally,

$$
\begin{aligned}
&= \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + \mathbb{E}(\mathbb{E}(\hat{\theta}) - \theta)^2 \\
&= \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + (\mathbb{E}(\hat{\theta}) - \theta)^2 \\
&= \mathsf{var}(\hat{\theta}) + \mathsf{bias}(\hat{\theta})^2
\end{aligned}
$$

Section 5

3

3a

Paraphrasing the problem, we assume that our data has a linear relationship. This means we have,

$$y = x^T\beta^* + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

So,

$$y|x \sim N(x^T\beta^*, \sigma^2)$$

We are asked to solve for $\hat{\beta}_{\text{MLE}}$ i.e find $\hat{\beta}_{\text{MLE}} = \arg\max_\beta L(\beta)$.

For multiple x_i, we typically write $y|X \sim N(X^T\beta^*, \sigma^2 I)$ for our input matrix X.

For multiple x_i, we typically write $y|X \sim N(X^T\beta^*, \sigma^2 I)$ for our input matrix X.

Our log likelihood will be,

$$\begin{aligned}\log L(\beta) &= \log P(y|X, \beta) \\ &= \log\left(\prod_{i=1}^{n} P(y_i|x_i, \beta)\right)\end{aligned}$$

For multiple x_i, we typically write $y|X \sim N(X^T\beta^*, \sigma^2 I)$ for our input matrix X.

Our log likelihood will be,

$$\begin{aligned}\log L(\beta) &= \log P(y|X, \beta) \\ &= \log\left(\prod_{i=1}^{n} P(y_i|x_i, \beta)\right) \\ &= \sum_{i=1}^{n} \log P(y_i|x_i, \beta)\end{aligned}$$

For multiple x_i, we typically write $y|X \sim N(X^T\beta^*, \sigma^2 I)$ for our input matrix X.

Our log likelihood will be,

$$\begin{aligned}
\log L(\beta) &= \log P(y|X, \beta) \\
&= \log\left(\prod_{i=1}^{n} P(y_i|x_i, \beta)\right) \\
&= \sum_{i=1}^{n} \log P(y_i|x_i, \beta) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}\right)\right)
\end{aligned}$$

$$= n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$

$$= n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$
$$= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|y - X\beta\|_2^2$$

So, to find $\hat{\beta}_{\mathsf{MLE}}$, we solve:

$$= n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$
$$= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|y - X\beta\|_2^2$$

So, to find $\hat{\beta}_{\mathsf{MLE}}$, we solve:

$$\hat{\beta}_{\mathsf{MLE}} = \arg\max_{\beta}\left(-\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right)$$
$$= \arg\max_{\beta} -\frac{1}{2\sigma^2}\|y - X\beta\|_2^2$$

$$= n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$
$$= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|y - X\beta\|_2^2$$

So, to find $\hat{\beta}_{\mathsf{MLE}}$, we solve:

$$\hat{\beta}_{\mathsf{MLE}} = \arg\max_{\beta}\left(-\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right)$$
$$= \arg\max_{\beta} -\frac{1}{2\sigma^2}\|y - X\beta\|_2^2$$
$$= \arg\min_{\beta} \|y - X\beta\|_2^2$$

This is just the least squares problem. So our solution is

$$\begin{aligned}\hat{\beta}_{\mathsf{MLE}} &= (X^T X)^{-1} X^T y \\ &= \hat{\beta}_{\mathsf{LS}}\end{aligned}$$

Section 6

Norms

Norms

We define the p-norm of a vector $x = (x_1, x_2, \ldots, x_n)$ as:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

For a norm g and vectors x, y, g needs to satisfy the following conditions to be a valid norm,

1. Triangle inequality. $g(x + y) \leq g(x) + g(y)$
2. Absolute homogeneity. For a constant c, $g(cx) = |c|g(x)$
3. Positive definiteness. The vector 0 should have norm 0.

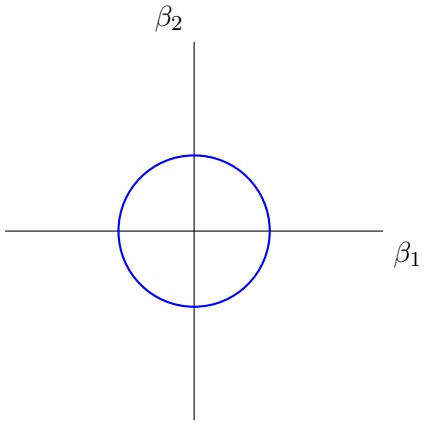
Euclidean norm

We've already encountered the Euclidean or $\ell 2$ norm as

$$\|x\|_2 = \sqrt{\sum_i x_i^2}$$

If we have a vector $\beta = (\beta_1, \beta_2)$ we can geometrically interpret the 2-norm as

$$\|x\|_2 = \sqrt{\beta_1^2 + \beta_2^2}$$

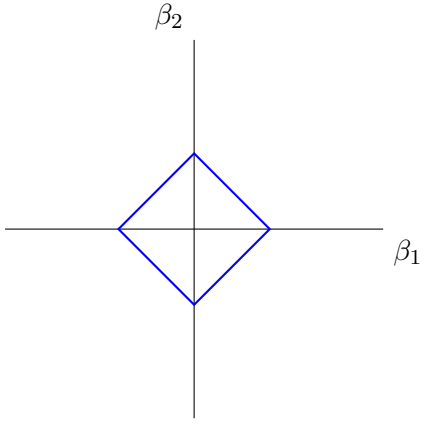


$\ell 1$ norm

The $\ell 1$ norm is defined as:

$$\|x\|_1 = \sum_i |x_i|$$

Again, if we have a vector $\beta = (\beta_1, \beta_2)$, the plot of the $\ell 1$ norm is:

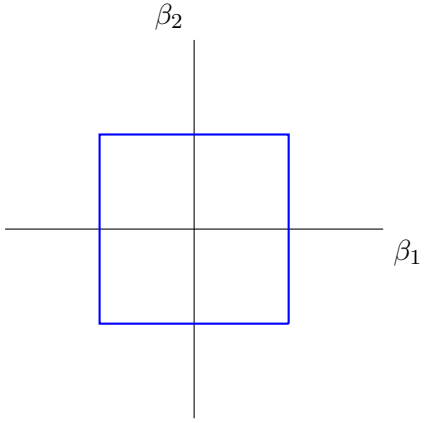


∞ norm

The ∞ norm is defined as:

$$\|x\|_\infty = \max_i |x_i|$$

Again, if we have a vector $\beta = (\beta_1, \beta_2)$, the plot of the ∞ norm is:



Section 7

3 (a, b, c)

3a

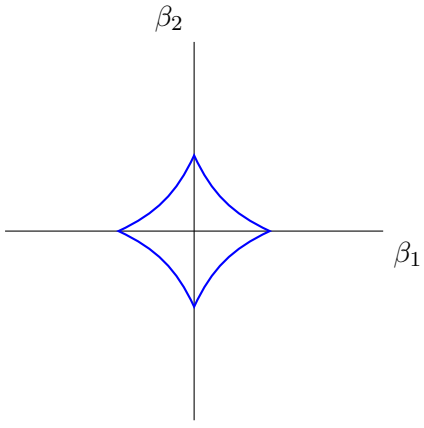
What is special about the $p = 0.5$ norm?

3a

What is special about the $p = 0.5$ norm?
The 0.5 norm is defined as:

$$\|x\|_{0.5} = \left(\sum_i \sqrt{x_i} \right)^2$$

Let's look further into this result, as we could possibly be breaking the rules of norms.



Take a point $x = (0, x_1)$ and $y = (y_1, 0)$,

$$\begin{aligned} \|x\|_{0.5} &= x_1 \\ \|y\|_{0.5} &= y_1 \\ \|x + y\|_{0.5} &= (\sqrt{x_1} + \sqrt{y_1})^2 = x_1 + 2\sqrt{x_1}\sqrt{y_1} + y_1 \end{aligned}$$

So, $\|x + y\|_{0.5} > \|x\|_{0.5} + \|y\|_{0.5}$ and the triangle inequality does not hold. Therefore, the $p = 0.5$ is not a valid norm.

3b

Describe the difference in the Ridge and LASSO problems explicitly.

3b

Describe the difference in the Ridge and LASSO problems explicitly.

The ridge regression problem can be written as:

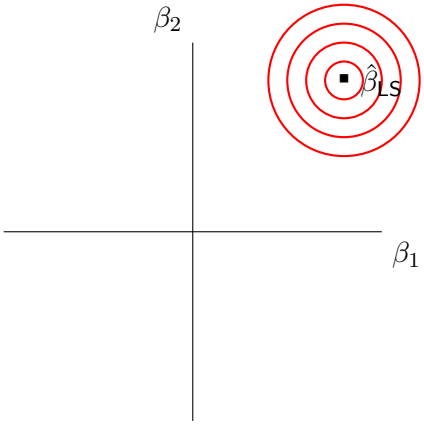
$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta}\{\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\}$$

We can interpret the term $\lambda\|\beta\|_2^2$ in the minimisation as finding the β with the minimum $2-$norm (multiplied by λ) while solving the least squares problem.

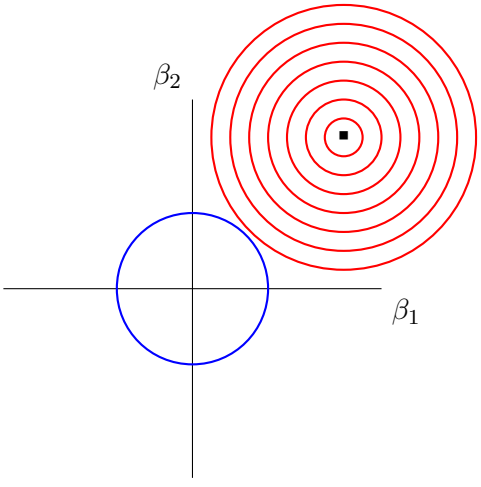
So, for an arbitrary k, we can redefine our problem as

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta}\{\|y - X\beta\|_2^2\} \qquad \text{where } \|\beta\|_2 \leq k$$

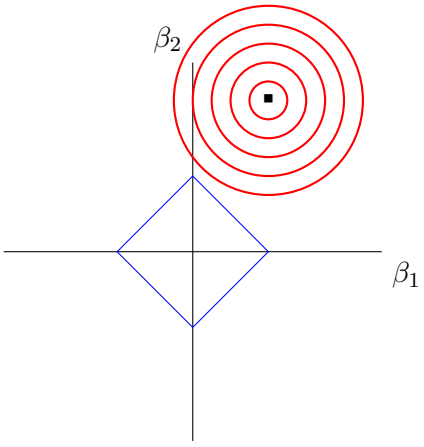
For the least squares setting, we have



For the ridge regression problem we add the constraint $\|\beta\|_2^2 \leq k$



For the lasso problem, we have the constraint $\|\beta\|_1 \leq k$. So,



3c

LASSO is said to induce sparsity. What does this mean? Why is it desirable?

3c

LASSO is said to induce sparsity. What does this mean? Why is it desirable?

We've seen from the plots that the $\ell 1$ norm tends to lie on the axes of the dimensions its applied in. This constraint will push smaller parameters towards zero and make the parameter vector *sparse* (i.e with fewer non-zero entries).

3c

LASSO is said to induce sparsity. What does this mean? Why is it desirable?

We've seen from the plots that the $\ell 1$ norm tends to lie on the axes of the dimensions its applied in. This constraint will push smaller parameters towards zero and make the parameter vector *sparse* (i.e with fewer non-zero entries).

This is especially useful when using a model for inference, where we can extract the *importance* or weighting of features clearer.