10-701 Machine Learning Recitation

Suyash Shringarpure

October 10, 2011

Outline of Topics

1. Homework/Lectures
2. Bias-Variance tradeoff
   - Theory and motivation
   - Bias and variance of KNN
3. VC dimension computation
4. Feature and model selection

Questions about HW 2 or the lectures?

- Project proposals due Monday 10/17
- Questions about Homework 2 or the lectures?
- Office Hours: Wednesday 2:00 pm-3:00 pm

Motivation

- Objective: Predict t from x using training set D of observations (x, t)

Motivation

- Objective: Predict t from x using training set D of observations (x, t)
- The (unknown) optimal predictor is $h(x) = \mathrm{E}_{\mathrm{D}}[t|x]$ (we won't prove this).

Motivation

- Objective: Predict t from x using training set D of observations (x, t)
- The (unknown) optimal predictor is $h(x) = \mathrm{E}_{\mathrm{D}}[t|x]$ (we won't prove this).
- Let your predictor be $y(x; \mathrm{D})$ based on your optimization method of choice.

Motivation

- Objective: Predict t from x using training set D of observations (x, t)
- The (unknown) optimal predictor is $h(x) = \mathrm{E}_{\mathrm{D}}[t|x]$ (we won't prove this).
- Let your predictor be $y(x; \mathrm{D})$ based on your optimization method of choice.
- Note that the predictor $y(x; \mathrm{D})$ depends on D.

- Objective: Predict t from x using training set D of observations (x, t)
- The (unknown) optimal predictor is $h(x) = \mathrm{E}_{\mathrm{D}}[t|x]$ (we won't prove this).
- Let your predictor be $y(x; \mathrm{D})$ based on your optimization method of choice.
- Note that the predictor $y(x; \mathrm{D})$ depends on D.
- What can we say about the error of $y(x; \mathrm{D})$?

Analysis

- Assume squared loss to measure the error of an predictor.
- Expected loss $= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - h(x))^2$.

Analysis

- Assume squared loss to measure the error of an predictor.
- Expected loss $= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - h(x))^2$.
- Note that the expectation is over D since your estimator $y(x;\mathrm{D})$ will be different for different D.

Analysis

- Assume squared loss to measure the error of an predictor.
- Expected loss $= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - h(x))^2$.
- Note that the expectation is over D since your estimator $y(x;\mathrm{D})$ will be different for different D.
- We will look at this error with respect to the error of the mean predictor $\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}))$.

- Assume squared loss to measure the error of an predictor.
- Expected loss $= \mathrm{E}_{\mathrm{D}}(y(x; \mathrm{D}) - h(x))^2$.
- Note that the expectation is over D since your estimator $y(x; \mathrm{D})$ will be different for different D.
- We will look at this error with respect to the error of the mean predictor $\mathrm{E}_{\mathrm{D}}(y(x; \mathrm{D}))$.
- Note that $\mathrm{E}_{\mathrm{D}}(y(x; \mathrm{D})) = \overline{y}(x)$, it is not a function of D anymore.

- $$\begin{aligned} \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - h(x))^2 &= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \overline{y}(x) + \overline{y}(x) - h(x))^2 \\ &= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \overline{y}(x))^2 + \mathrm{E}_{\mathrm{D}}(\overline{y}(x) - h(x))^2 \\ &\quad + \mathrm{E}_{\mathrm{D}}(2(y(x;\mathrm{D}) - \overline{y}(x))(\overline{y}(x) - h(x))) \end{aligned}$$

- $$\begin{aligned} \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - h(x))^2 &= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \overline{y}(x) + \overline{y}(x) - h(x))^2 \\ &= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \overline{y}(x))^2 + \mathrm{E}_{\mathrm{D}}(\overline{y}(x) - h(x))^2 \\ &\quad + \mathrm{E}_{\mathrm{D}}(2(y(x;\mathrm{D}) - \overline{y}(x))(\overline{y}(x) - h(x))) \end{aligned}$$

- Now, since $(\overline{y}(x) - h(x))$ is constant w.r.t D, and we know that $\mathrm{E}[cX] = c\mathrm{E}[X]$, we can simplify the last term $\mathrm{E}_{\mathrm{D}}(2(y(x;\mathrm{D}) - \overline{y}(x))(\overline{y}(x) - h(x)))$

- $$\begin{aligned} \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - h(x))^2 &= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \overline{y}(x) + \overline{y}(x) - h(x))^2 \\ &= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \overline{y}(x))^2 + \mathrm{E}_{\mathrm{D}}(\overline{y}(x) - h(x))^2 \\ &\quad + \mathrm{E}_{\mathrm{D}}(2(y(x;\mathrm{D}) - \overline{y}(x))(\overline{y}(x) - h(x))) \end{aligned}$$
- Now, since $(\overline{y}(x) - h(x))$ is constant w.r.t D, and we know that $\mathrm{E}[cX] = c\mathrm{E}[X]$, we can simplify the last term $\mathrm{E}_{\mathrm{D}}(2(y(x;\mathrm{D}) - \overline{y}(x))(\overline{y}(x) - h(x)))$
- $$\begin{aligned} \text{Last term} &= \mathrm{E}_{\mathrm{D}}(2(y(x;\mathrm{D}) - \overline{y}(x))(\overline{y}(x) - h(x))) \\ &= 2(\overline{y}(x) - h(x))\mathrm{E}_{\mathrm{D}}((y(x;\mathrm{D}) - \overline{y}(x))) \\ &= 2(\overline{y}(x) - h(x))(\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \overline{y}(x))) \\ &= 2(\overline{y}(x) - h(x))(\overline{y}(x)) - \overline{y}(x))) \\ &= 0 \end{aligned}$$

$$\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - h(x))^2 = \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \overline{y}(x))^2 + \mathrm{E}_{\mathrm{D}}(\overline{y}(x) - h(x))^2$$
$$= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})))^2 + \mathrm{E}_{\mathrm{D}}(\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})) - h(x))^2$$
$$= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})))^2 + (\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})) - h(x))^2$$

- Bias: $\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})) - h(x)$ - How far is the mean predictor from the optimal?

Bias and variance

$$\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - h(x))^2 = \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \overline{y}(x))^2 + \mathrm{E}_{\mathrm{D}}(\overline{y}(x) - h(x))^2$$
$$= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})))^2 + \mathrm{E}_{\mathrm{D}}(\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})) - h(x))^2$$
$$= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})))^2 + (\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})) - h(x))^2$$

- Bias: $\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})) - h(x)$ - How far is the mean predictor from the optimal?
- Variance: $\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}))^2$ - How far away is a given predictor from the mean predictor?

Bias and variance

$$\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - h(x))^2 = \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \bar{y}(x))^2 + \mathrm{E}_{\mathrm{D}}(\bar{y}(x) - h(x))^2$$
$$= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})))^2 + \mathrm{E}_{\mathrm{D}}(\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})) - h(x))^2$$
$$= \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})))^2 + (\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})) - h(x))^2$$

- Bias: $\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D})) - h(x)$ - How far is the mean predictor from the optimal?
- Variance: $\mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}) - \mathrm{E}_{\mathrm{D}}(y(x;\mathrm{D}))^2$ - How far away is a given predictor from the mean predictor?
- Expected loss = Variance + $(\text{Bias})^2$

$$E_D(y(x;D) - h(x))^2 = E_D(y(x;D) - \bar{y}(x))^2 + E_D(\bar{y}(x) - h(x))^2$$
$$= E_D(y(x;D) - E_D(y(x;D)))^2 + E_D(E_D(y(x;D)) - h(x))^2$$
$$= E_D(y(x;D) - E_D(y(x;D)))^2 + (E_D(y(x;D)) - h(x))^2$$

- Bias: $E_D(y(x;D)) - h(x)$ - How far is the mean predictor from the optimal?
- Variance: $E_D(y(x;D) - E_D(y(x;D))^2$ - How far away is a given predictor from the mean predictor?
- Expected loss = Variance + $(\text{Bias})^2$
- A more general analysis would account for the fact that your optimal predictor may itself not be perfect if there is noise in the model. Say $t = f(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$.

$$\mathrm{E_D}(y(x;\mathrm{D}) - h(x))^2 = \mathrm{E_D}(y(x;\mathrm{D}) - \overline{y}(x))^2 + \mathrm{E_D}(\overline{y}(x) - h(x))^2$$
$$= \mathrm{E_D}(y(x;\mathrm{D}) - \mathrm{E_D}(y(x;\mathrm{D})))^2 + \mathrm{E_D}(\mathrm{E_D}(y(x;\mathrm{D})) - h(x))^2$$
$$= \mathrm{E_D}(y(x;\mathrm{D}) - \mathrm{E_D}(y(x;\mathrm{D})))^2 + (\mathrm{E_D}(y(x;\mathrm{D})) - h(x))^2$$

- Bias: $\mathrm{E_D}(y(x;\mathrm{D})) - h(x)$ - How far is the mean predictor from the optimal?
- Variance: $\mathrm{E_D}(y(x;\mathrm{D}) - \mathrm{E_D}(y(x;\mathrm{D}))^2$ - How far away is a given predictor from the mean predictor?
- Expected loss = Variance + (Bias)^2
- A more general analysis would account for the fact that your optimal predictor may itself not be perfect if there is noise in the model. Say $t = f(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$.
- Then, Expected loss = Variance + (Bias)^2 + Noise (irreducible)

- Bias can be minimized by choosing the predictor from a complex family which can fit any data distribution well. This could cause overfitting.

Bias-Variance tradeoff

- Bias can be minimized by choosing the predictor from a complex family which can fit any data distribution well. This could cause overfitting.
- Variance can be minimized by choosing the predictor from a simple family so that all predictors are close to the mean. This could lead to a poor fit to the data.

- Bias can be minimized by choosing the predictor from a complex family which can fit any data distribution well. This could cause overfitting.
- Variance can be minimized by choosing the predictor from a simple family so that all predictors are close to the mean. This could lead to a poor fit to the data.
- The two minimization objectives are contradictory.

- Bias can be minimized by choosing the predictor from a complex family which can fit any data distribution well. This could cause overfitting.
- Variance can be minimized by choosing the predictor from a simple family so that all predictors are close to the mean. This could lead to a poor fit to the data.
- The two minimization objectives are contradictory.
- The solution is usually to settle for a predictor with intermediate values of bias and variance for the best generalization.

Alternative interpretations - Bias

- Bias: What is the difference between the true predictor and the predictor that you can expect to learn?

Alternative interpretations - Bias

- Bias: What is the difference between the true predictor and the predictor that you can expect to learn?
- It measures the impact of the assumptions you made when choosing your predictor class. How different are they from the reality of the data?

Alternative interpretations - Bias

- Bias: What is the difference between the true predictor and the predictor that you can expect to learn?
- It measures the impact of the assumptions you made when choosing your predictor class. How different are they from the reality of the data?
- For example, suppose I'm using logistic regression (a linear classifier).

Alternative interpretations - Bias

- Bias: What is the difference between the true predictor and the predictor that you can expect to learn?
- It measures the impact of the assumptions you made when choosing your predictor class. How different are they from the reality of the data?
- For example, suppose I'm using logistic regression (a linear classifier).
- Qualitatively, is the bias more when - (a) the decision boundary is non-linear (b) the decision boundary is linear?

Alternative interpretations - Bias

- Bias: What is the difference between the true predictor and the predictor that you can expect to learn?
- It measures the impact of the assumptions you made when choosing your predictor class. How different are they from the reality of the data?
- For example, suppose I'm using logistic regression (a linear classifier).
- Qualitatively, is the bias more when - (a) the decision boundary is non-linear (b) the decision boundary is linear?
- When the decision boundary is non-linear, I can never expect to perfectly learn the decision boundary. So the bias is higher.

Alternative interpretations - Variance

- Variance can be thought of as the sensitivity of the predictor to the dataset D.

Alternative interpretations - Variance

- Variance can be thought of as the sensitivity of the predictor to the dataset D.
- A very complex (flexible) predictor will change a lot even if the dataset D changes due to just noise.

Alternative interpretations - Variance

- Variance can be thought of as the sensitivity of the predictor to the dataset D.
- A very complex (flexible) predictor will change a lot even if the dataset D changes due to just noise.
- Sensitivity to noise is undesirable.

Prediction with KNN

- Suppose we want to use a KNN to predict $t = f(x) + \epsilon$ ($\epsilon =$ noise). We are given pairs of the form $x, t(x)$.

- Suppose we want to use a KNN to predict $t = f(x) + \epsilon$ ($\epsilon=$ noise). We are given pairs of the form $x, t(x)$.
- For prediction, if we use k nearest neighbors, the prediction rule is

$$t(x) = \sum_{y : y \in N_k(x)} \frac{t(y)}{k}$$

where $N_k(x)$ is the set of the k nearest neighbors of point x.

- Suppose we want to use a KNN to predict $t = f(x) + \epsilon$ ($\epsilon =$ noise). We are given pairs of the form $x, t(x)$.
- For prediction, if we use k nearest neighbors, the prediction rule is

$$t(x) = \sum_{y : y \in N_k(x)} \frac{t(y)}{k}$$

where $N_k(x)$ is the set of the k nearest neighbors of point x.
- First, what does the bias and variance of this predictor depend on?

Prediction with KNN

- Suppose we want to use a KNN to predict $t = f(x) + \epsilon$ ($\epsilon =$ noise). We are given pairs of the form $x, t(x)$.
- For prediction, if we use k nearest neighbors, the prediction rule is

$$t(x) = \sum_{y : y \in N_k(x)} \frac{t(y)}{k}$$

  where $N_k(x)$ is the set of the k nearest neighbors of point x.
- First, what does the bias and variance of this predictor depend on?
- It depends on k, the number of neighbors.

- What if $k = 1$?

Bias-variance and the number of neighbors k

- What if $k = 1$?
- Then $t(x)$ is assumed to be equal to the value for its nearest neighbor.

Bias-variance and the number of neighbors k

- What if $k = 1$?
- Then $t(x)$ is assumed to be equal to the value for its nearest neighbor.
- Assuming f is smooth, $t(x)$ should be close to the value for its neighbor. So the bias will be low.

Bias-variance and the number of neighbors k

- What if $k = 1$?
- Then $t(x)$ is assumed to be equal to the value for its nearest neighbor.
- Assuming f is smooth, $t(x)$ should be close to the value for its neighbor. So the bias will be low.
- But if the dataset changes even a little, the nearest neighbor for x could change. So the prediction could be quite different. So the variance is high.

Bias-variance and the number of neighbors k

- What if $k = 1$?
- Then $t(x)$ is assumed to be equal to the value for its nearest neighbor.
- Assuming f is smooth, $t(x)$ should be close to the value for its neighbor. So the bias will be low.
- But if the dataset changes even a little, the nearest neighbor for x could change. So the prediction could be quite different. So the variance is high.
- Suppose $k = N$, the number of training points.

Bias-variance and the number of neighbors k

- What if $k = 1$?
- Then $t(x)$ is assumed to be equal to the value for its nearest neighbor.
- Assuming f is smooth, $t(x)$ should be close to the value for its neighbor. So the bias will be low.
- But if the dataset changes even a little, the nearest neighbor for x could change. So the prediction could be quite different. So the variance is high.
- Suppose $k = N$, the number of training points.
- The prediction $t(x)$ is then the mean t of all N training points. It is completely independent of what point x you want to make the prediction for.

Bias-variance and the number of neighbors k

- What if $k = 1$?
- Then $t(x)$ is assumed to be equal to the value for its nearest neighbor.
- Assuming f is smooth, $t(x)$ should be close to the value for its neighbor. So the bias will be low.
- But if the dataset changes even a little, the nearest neighbor for x could change. So the prediction could be quite different. So the variance is high.
- Suppose $k = N$, the number of training points.
- The prediction $t(x)$ is then the mean t of all N training points. It is completely independent of what point x you want to make the prediction for.
- Therefore, the bias is high and the variance is low.

KNN bias-variance summary

- Bias $\propto$ Number of nearest neighbors.
- Variance $\propto \frac{1}{\text{Number of nearest neighbors}}$

- Bias $\propto$ Number of nearest neighbors.
- Variance $\propto \frac{1}{\text{Number of nearest neighbors}}$
- Earlier, we said that a more complex predictor has less bias and more variance.

KNN bias-variance summary

- Bias $\propto$ Number of nearest neighbors.
- Variance $\propto \frac{1}{\text{Number of nearest neighbors}}$
- Earlier, we said that a more complex predictor has less bias and more variance.
- This suggests that KNN complexity actually reduces as k increases. (Not intuitive!)

Computing VC dimension of a classifer

- A classifer family shatters a set of points if for any labeling of the set of points, there exists a member of the classifer family that can correctly label the set.

Computing VC dimension of a classifer

- A classifer family shatters a set of points if for any labeling of the set of points, there exists a member of the classifer family that can correctly label the set.
- If a classifier family has VC dimension at least m, then there must be a set of m points it can shatter.

Computing VC dimension of a classifer

- A classifer family shatters a set of points if for any labeling of the set of points, there exists a member of the classifer family that can correctly label the set.
- If a classifier family has VC dimension at least m, then there must be a set of m points it can shatter.
- If a classifier family has VC dimension at most m, then there cannot be any set of $m + 1$ points that it can shatter.

Computing VC dimension of a classifer

- A classifer family shatters a set of points if for any labeling of the set of points, there exists a member of the classifer family that can correctly label the set.
- If a classifier family has VC dimension at least m, then there must be a set of m points it can shatter.
- If a classifier family has VC dimension at most m, then there cannot be any set of $m+1$ points that it can shatter.
- A classifier family has VC dimension m if its VC dimension is at least m and at most m.

Open-intervals (in one direction):
H1: if $x > a$ then $y = 1$ else $y = 0$.

Open-intervals (in both directions):
H2: if $x > a$ then $y = 1$ else $y = 0$
or if $x < a$ then $y = 1$ else $y = 0$

H3: if $a < x < b$ then $y = 1$ else $y = 0$.

H4: if $a < x < b$ then $y = 1$ else $y = 0$
or if $a < x < b$ then $y = 0$ else $y = 1$

Feature and model selection- linear regression

- Setting: Data: X$= 50 \times 15$, Y $= 50 \times 1$. 50 data points, 15 features.

Feature and model selection- linear regression

- Setting: Data: $X = 50 \times 15$, $Y = 50 \times 1$. 50 data points, 15 features.
- Assume a linear regression model. You can choose to include or exclude features.

Feature and model selection- linear regression

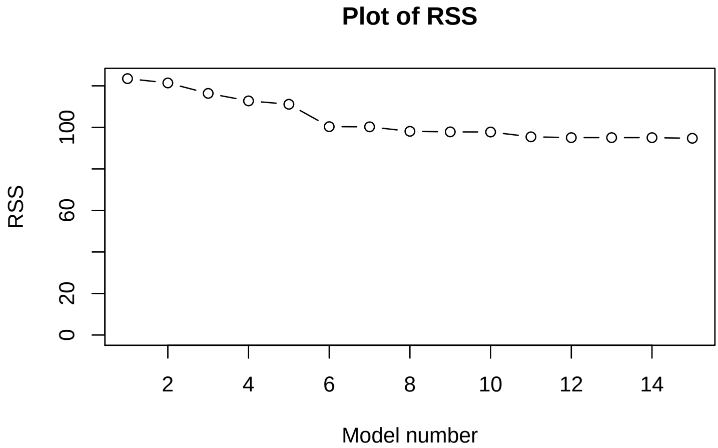
- Setting: Data: X= 50×15 , Y $= 50 \times 1$. 50 data points, 15 features.
- Assume a linear regression model. You can choose to include or exclude features.
- What I did: Built models by progressively adding features. Model M_i used all 50 data points and the first i features, i.e X(M_i)=X[1:50,1:i].

Feature and model selection- linear regression

- Setting: Data: X= 50×15 , Y $= 50 \times 1$. 50 data points, 15 features.
- Assume a linear regression model. You can choose to include or exclude features.
- What I did: Built models by progressively adding features. Model M_i used all 50 data points and the first i features, i.e X(M_i)=X[1:50,1:i].
- Fit a linear regression model using the *lm* function in R.

Feature and model selection- linear regression

- Setting: Data: X= 50×15 , Y = 50×1. 50 data points, 15 features.
- Assume a linear regression model. You can choose to include or exclude features.
- What I did: Built models by progressively adding features. Model M_i used all 50 data points and the first i features, i.e X(M_i)=X[1:50,1:i].
- Fit a linear regression model using the *lm* function in R.
- Question: Which model $(M_1, \cdots, M_{15})$ is best?



What model is best?

Aside about RSS and extra features

- The graph showed RSS decreasing as more features were added.

Aside about RSS and extra features

- The graph showed RSS decreasing as more features were added.
- Claim: The performance of linear regression can never get worse due to addition of a feature (assuming no optimization errors).

Aside about RSS and extra features

- The graph showed RSS decreasing as more features were added.
- Claim: The performance of linear regression can never get worse due to addition of a feature (assuming no optimization errors).
- Proof?

Aside about RSS and extra features

- The graph showed RSS decreasing as more features were added.
- Claim: The performance of linear regression can never get worse due to addition of a feature (assuming no optimization errors).
- Proof?
- Useful to note that there are no direct interactions between features in linear regression.

Back to model/feature selection

- We will try to use information criteria to resolve this problem.

Back to model/feature selection

- We will try to use information criteria to resolve this problem.
- Remember the gaussian model equivalent of linear regression?

Back to model/feature selection

- We will try to use information criteria to resolve this problem.
- Remember the gaussian model equivalent of linear regression?
- $y_i = \theta^T x_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Back to model/feature selection

- We will try to use information criteria to resolve this problem.
- Remember the gaussian model equivalent of linear regression?
- $y_i = \theta^T x_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- Log-likelihood $l(\theta) = n \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$

Back to model/feature selection

- We will try to use information criteria to resolve this problem.
- Remember the gaussian model equivalent of linear regression?
- $y_i = \theta^T x_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- Log-likelihood $l(\theta) = n \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$
- Here, $\sigma^2 = 0.5$(assume known), so $l(\theta) = -\sum_{i=1}^{n} (y_i - \theta^T x_i)^2$ $+$ C

Bayesian Information Criterion

- $BIC = l(\theta) - \frac{k}{2}\log n$, where k is the number of free parameters

- $BIC = l(\theta) - \frac{k}{2} \log n$, where k is the number of free parameters
- For linear regression with i features, there are $i + 1$ free parameters.

Bayesian Information Criterion

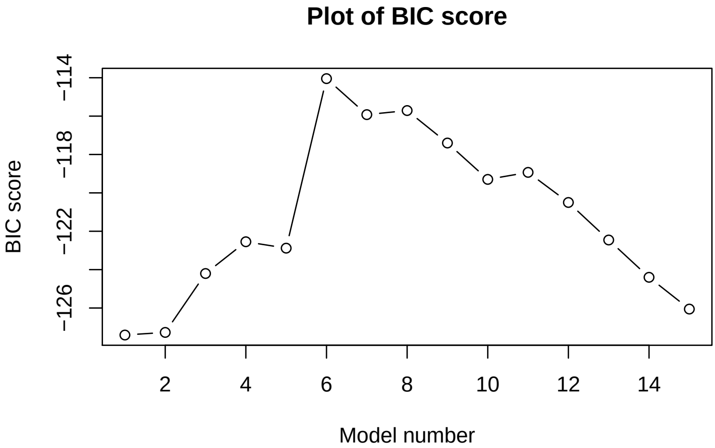
- $BIC = l(\theta) - \frac{k}{2}\log n$, where k is the number of free parameters
- For linear regression with i features, there are $i + 1$ free parameters.
- $BIC(M_i) = -\sum_{i=1}^{n}(y_i - \theta^T x_i)^2 - \frac{i+1}{2}\log n$

Bayesian Information Criterion

- $BIC = l(\theta) - \frac{k}{2}\log n$, where k is the number of free parameters
- For linear regression with i features, there are $i+1$ free parameters.
- $BIC(M_i) = -\sum_{i=1}^{n}(y_i - \theta^T x_i)^2 - \frac{i+1}{2}\log n$
- The model with maximum BIC is considered the best.

Bayesian Information Criterion

- $BIC = l(\theta) - \frac{k}{2} \log n$, where k is the number of free parameters
- For linear regression with i features, there are $i + 1$ free parameters.
- $BIC(M_i) = -\sum_{i=1}^{n}(y_i - \theta^T x_i)^2 - \frac{i+1}{2} \log n$
- The model with maximum BIC is considered the best.
- (Note: Alternative definition of BIC possible)



So the best model is ?