STAT 535

10/20/22

Lecture 7

Bias-Variance tradeoff

Lecture II: Prediction – Basic concepts

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

October, 2022

Parametric vs non-parametric ✓

Generative and discriminative models for classification ✓

- Generative classifiers
- Discriminative classifiers
- Generative vs discriminative classifiers

Loss functions ✓

- Bayes loss ~

Variance, bias and complexity ←

Reading HTF Ch.: 2.1–5,2.9, 7.1–4 bias-variance tradeoff, Murphy Ch.: 1., 8.6[1], Bach Ch.:

[1] Neither textbook is close to these notes except in a few places; take them as alternative perspectives or related reading

The "learning" problem

- **Given**
- a problem (e.g. recognize digits from $m \times m$ gray-scale images)
- a **sample** or (**training set**) of **labeled data**

$$\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$$

$\mathcal{D}_n$

drawn i.i.d. from an unknown P_{XY}

- **model class** $\mathcal{F} = \{f\}$ = set of predictors to choose from

- **Wanted**
- a predictor $f \in \mathcal{F}$ that performs well on future samples from the same P_{XY}

  - "choose a predictor $f \in \mathcal{F}$" = training/learning
  - "performs well on future samples" (i.e. f **generalizes** well) – how do we measure this? how can we "guarantee" it?
  - choosing $\mathcal{F}$ is the **model selection problem** – about this later

$\mathcal{D}_n \sim P_{XY}^n$

P_{XY}

$\mathcal{F}$

n

learning algorithm

actors in Bias-variance t-off

A zoo of predictors

- Linear regression
- Logistic regression
- Linear Discriminant (LDA)
- Quadratic Discriminant (QDA)
- CART (Decision Trees)
- K-Nearest Neighbors
- Nadaraya-Watson (Kernel regression)
- Naive Bayes
- Neural networks/Deep learning
- Support Vector Machines
- Monotonic Regression

Bayes loss

- How small can the expected loss $L(f)$ be?
  It is clear that
  $$L(\mathcal{F}) = \min_{f \in \mathcal{F}} L(f) \geq \min_{f} L(f) = L^* \tag{10}$$
  where L^* is taken over all possible functions f that take values in $\mathcal{Y}$.
- L^* is the absolute minimum loss for the given P_{XY} and it is called the **Bayes loss**.
- The Bayes loss is usually not zero

Bayes loss for (binary) classification

- Fix x and assume $P_{Y|X}$ known. Then:
  - Label y will have probability $P_{Y|X}(y|x)$ at this x.
  - No deterministic guess $f(x)$ for y will make the classification error $E_{P_{Y|X=x}}[L_{01}(y, f(x))]$ (unless $P_{Y|X=x}$ is itself deterministic)
  - Best guess minimizes the probability of being wrong. This is achieved by chosing the most probable class

$$y^*(x) = \operatorname{argmax}_y P_{Y|X}(y|x) \tag{11}$$

  - The probability of being wrong if we choose $y^*(x)$ is $1 - p^*(x)$, where $p^*(x) = \max_y P_{Y|X}(y|x)$.
- The **Bayes classifier** is $y^*(x)$ as a function of x and its expected loss is the Bayes loss

$$L_{01}^* = E_{P_X}[1 - p^*(X)] = E_{P_X}[1 - \max_y P[Y|X]] \tag{12}$$

This shows that the Bayes loss is a property of the problem, via L and P_{XY}, and not of any model class or learning algorithm.

Example

In a classification problem where the class label depends deterministically of the input, the Bayes loss is 0. For example, classifying between written English and written Japanese has (probably) zero Bayes loss.

Example

Consider the least squares loss and the following data distribution: $P_{Y|X} \sim N(g(X), \sigma^2)$. In other words, the Y values are normally distributed around a deterministic function $g(X)$. In this case, optimal least squares predictor is the mean of Y given X, which is equal to $g(X)$. The Bayes loss is the expected squared error around the mean, which is σ^2. Exercise what is the expression of the Bayes loss if $P_{Y|X} \sim N(g(X), \sigma(X)^2)$?

Exercise What is the Bayes loss if (1) $P(Y|X) \sim N((\beta^*)^T X, \sigma^2 I)$ and the loss is L_{LS}; (2) $P(X|Y = \pm 1) \sim N(\mu_\pm, \sigma^2 I)$ and the loss is L_{01} (for simplicity, assume $X \in \mathbb{R}, \mu_{pm} = \pm 1, \sigma = 1$); (3) give a formula for the Bayes loss if we know $P(X|Y = \pm 1), P(Y)$, $Y \in \{\pm 1\}$ and the loss is L_{01}. (4) Give an example of a situation when the Bayes loss is 0.

Bias and variance: definitions (never to be used again)

Preliminaries

- What we have a data source P_{XY} and a class of predictors $\mathcal{F}$
- From P_{XY} we sample i.i.d. $\mathcal{D}_N$ of size n. Hence $\mathcal{D}_N \sim P_{XY}^n$. []

Bias and Variance as in Intro Stat Theory

- We want to estimate a parameter $\theta \in \Theta \subseteq \mathbb{R}$
- We use $\mathcal{D}_N$ to obtain estimator $\hat{\theta}_{\mathcal{D}_N}$ which is a function of $\mathcal{D}_N$.
- $\mathcal{D}_N$ is random, hence so is $\hat{\theta}_{\mathcal{D}_N}$
- Bias $= E_{P^n}[\hat{\theta}_{\mathcal{D}_N}] - \theta$
- Variance $= Var_{P^n}(\hat{\theta}_{\mathcal{D}_N})$

  Both Bias and Variance are computed under the distribution from which we sampled $\mathcal{D}_N$, denoted by P^n.

Bias and Variance for us

- We use $\mathcal{D}_N$ to estimate $\hat{f}_N \in \mathcal{F}$

$$\hat{f}_{\mathcal{D}_N} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{L}(f, \mathcal{D}_N) \tag{15}$$

- $\mathcal{D}_N$ is random, hence so if $\hat{f}_N$.

- Main differences
  1. $\hat{f}$ is a function!
  2. We are interested in the predictions and not the parameters of $\hat{f}$.
- Several proposals to define bias and variance exist.
- Bias and variance are properties of $\mathcal{F}$.

What we need to know in this course is qualitative

Bias & Var in Prediction

$P_{XY}^n \longrightarrow \mathcal{D}_n$ (random) $\longrightarrow$ Learning alg $\longrightarrow \hat{f}_{\mathcal{D}_n}$ (random) $: \mathbb{R}^d \to \mathcal{Y}$ (output) $\longrightarrow L(\hat{f}_{\mathcal{D}_n}) = \mathbb{E}_{P_{XY}}[L(Y, \hat{f}_{\mathcal{D}_n}(X))]$ (random)

ML Algo Theory

$\mathcal{F}$

$\mathcal{Y}$: $\mathbb{R}$, labels, ...

$\mathbb{E}_{P_{XY}^n}[\hat{f}_{\mathcal{D}_n}] - f^{true}$?? $\longrightarrow$ No f^{true} !!

No $\mathbb{E}[\ ]$ on $\mathcal{Y}$ when $\mathcal{Y} \not\subset \mathbb{R}$

- Case I $\quad Y = f^{true}(X) + \varepsilon \quad \varepsilon$ iid $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Var}\,\varepsilon = \sigma^2 \quad Y, f(X) \in \mathbb{R}$

typical for L_{LS}, Regression

$$\mathbb{E}_{P_{XY}^n}\left[\left(\hat{f}_{\mathcal{D}_n} - f^{true}\right)^2\right] = \mathbb{E}_{P_{XY}^n}\left[\left(\hat{f}_{\mathcal{D}_n} - \mathbb{E}_{P_{XY}^n}[\hat{f}_{\mathcal{D}_n}]\right)^2 + \left(\mathbb{E}_{P_{XY}^n}[\hat{f}_{\mathcal{D}_n}] - f^{true}\right)^2 - 2(\ )(\ )\right]$$

$\mathbb{E}_{P_{XY}^n} = 0$

at $x \in \mathbb{R}$: L_{LS} at x

integrated over x: $\mathbb{E}_{P_{XY}^n}[L_{LS}(\hat{f}_{\mathcal{D}_n})]$

fction of x

$$= \mathbb{E}_{P_{XY}^n}\left[\left(\hat{f}_{\mathcal{D}_n} - \mathbb{E}_{P_{XY}^n}[\hat{f}_{\mathcal{D}_n}]\right)^2\right] + \left(\mathbb{E}_{P_{XY}^n}[\hat{f}_{\mathcal{D}_n}] - f^{true}\right)^2$$

$\mathrm{Var}\,\hat{f}_{\mathcal{D}_n}$ $\qquad$ Bias^2

smoother $\Rightarrow$ "easier to fit"

Bias–Var decomposition of L_{LS}

Lecture Notes II.1 – Bias and variance in Kernel Regression

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

October, 2021

An elementary analysis

Bias, Variance and h for $x \in \mathbb{R}$

Kernel regression by Nadaraya-Watson

$$\hat{f}_{\mathcal{D}_n} \equiv \hat{y}(x) = \frac{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right) y^i}{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right)} \tag{1}$$

Let $w_i = \dfrac{b\left(\frac{||x-x^i||}{h}\right)}{\sum_{i'=1}^{n} b\left(\frac{||x-x^{i'}||}{h}\right)}$.

Assumptions

A0 For simplicity, in this analysis we assume $x \in \mathbb{R}$.

A1 There is a true smooth[1] function $f(x)$ so that

$$y = f(x) + \varepsilon, \tag{2}$$

where ε is sampled independently for each x from a distribution P_ε, with $E_{P_\varepsilon}[\varepsilon] = 0$, $Var_{P_\varepsilon}(\varepsilon) = \sigma^2$.

A2 The kernel $b(z)$ is smooth, $\int_{\mathbb{R}} b(z)dz = 1$, $\int_{\mathbb{R}} zb(z) = 0$, and we denote $\sigma_b^2 = \int_{\mathbb{R}} z^2 b(z)dz$, $\gamma_b^2 = \int_{\mathbb{R}} b^2(z)dz$.

In this first analysis, we consider that the values x, $x^{1:N}$ are fixed; hence, the randomness is only in $\varepsilon^{1:N}$.

fixed ↗ single $\mathcal{D}_n$

[1] with continuous derivatives up to order 2

Expectation of $\hat{y}(x)$ – a simple analysis

Expanding f in Taylor series around x we obtain

$$f(x^i) = f(x) + f'(x)(x^i - x) + \frac{f''(x)}{2}(x^i - x)^2 + o((x^i - x)^2) \tag{3}$$

We also have

$$y^i = f(x^i) + \varepsilon^i. \tag{4}$$

We now write the expectation of $\hat{y}(x)$ from (1), replacing in it y^i and $f(x^i)$ as above. What we would like to happen is that this expectation equals $f(x)$. Let us see if this is the case.

$$E_{P_\varepsilon^n}[\hat{y}(x)] = E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i y^i\right] = E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i \left(f(x^i) + \varepsilon^i\right)\right] \tag{5}$$

$$= \sum_{i=1}^n w_i f(x) + \sum_{i=1}^n w_i f'(x)(x^i - x) + \sum_{i=1}^n w_i \frac{f''(x)}{2}(x^i - x)^2 + \underbrace{E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i \varepsilon^i\right]}_{=0]} \tag{6}$$

$$= f(x) + \underbrace{f'(x)\sum_{i=1}^n w_i (x^i - x) + \frac{f''(x)}{2}\sum_{i=1}^n w_i (x^i - x)^2}_{\text{bias}} \tag{7}$$

In the above, the expressions in red depend of f, those in blue depend on x and $x^{1:N}$.

Qualitative analysis of the bias terms

The first order term $f'(x)\sum_{i=1}^{n} w_i(x^i - x)$ is responsible for **border effects**.
The second order term **smooths out** sharp peaks and valleys.

$= 0$?

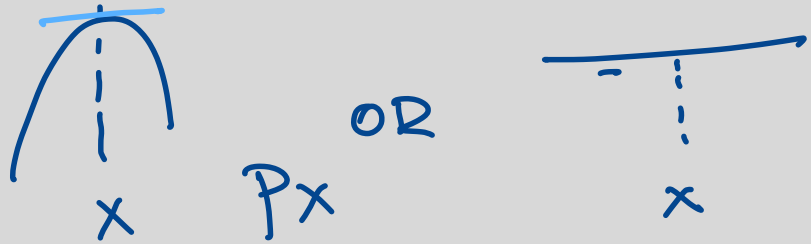
if $f'(x) = 0$

if $\sum_{i=1}^{n} w_i (x^i - x) = \sum_i w_i x^i - \left(\sum w_i\right) x = 0$

$\sum_i w_i x^i$: weighted avg

$\sum w_i$: 1

$= x \Leftrightarrow$ neighbors are symmetric around x



$\Rightarrow \nabla p_x(x) = 0$

p_x: density of x^i

Qualitative analysis of the bias terms

The first order term ~~$f'(x)\sum_{i=1}^{n} w_i(x^i - x)$~~ is responsible for **border effects**.
The second order term **smooths out** sharp peaks and valleys.

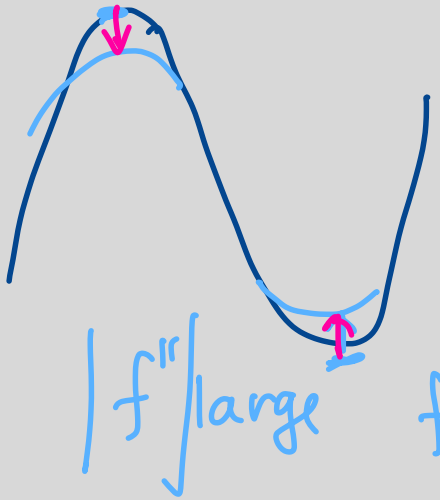
$f''(x) \sum_i w_i (x^i - x)^2 = 0$ if

$f''(x) = 0$ OR ~~$= 0$~~

weighted square distance > 0

reduce by $h \downarrow$, dense sampling

large for f



Bias, Variance and h for $x \in \mathbb{R}$

2

The **bias** of $\hat{y}$ at x is defined as $E_{P_X^n} E_{P_\varepsilon^n} [\hat{y}(x) - f(x)]$.

$$E_{P_X^n} E_{P_\varepsilon^n} [\hat{y}(x) - f(x)] = h^2 \sigma_b^2 \left(\frac{f'(x) p_X'(x)}{p_X(x)} + \frac{f''(x)}{2} \right) + o(h^2) \tag{8}$$

The **variance** $\hat{y}$ at x is defined as $Var_{P_X^n P_\varepsilon^n}(\hat{y}(x))$.

$$Var_{P_X^n} P_\varepsilon^n(\hat{y}(x)) = \frac{\gamma^2}{nh} \sigma^2 + o\left(\frac{1}{nh}\right). \tag{9}$$

The **MSE (Mean Squared Error)** is defined as $E_{P_X^n} E_{P_\varepsilon^n} \left[(\hat{y}(x) - f(x))^2 \right]$, which equals

$$MSE(x) = \textbf{bias}^2 + \textbf{variance}^2 = h^4 \sigma_b^4 \left(\frac{f'(x) p_X'(x)}{p_X(x)} + \frac{f''(x)}{2} \right) + \frac{\gamma_b^2}{nh} \sigma^2 + \dots \tag{10}$$

[2] After []

Optimal selection of h

If the MSE is integrated over $\mathbb{R}$ we obtain the **MISE**$= \int_{\mathbb{R}} MSE(x)dx$.
The kernel width h can be chosen to minimize the MISE, for fixed f, p_X and b.
We set to 0 the partial derivative

$$\frac{\partial MISE}{\partial h} = h^3 \left(\blacksquare \right) - \frac{(\blacksquare)}{nh^2} = 0. \tag{11}$$

It follows that $h^5 \propto \frac{1}{n}$, or

$$h \propto \frac{1}{N^{1/5}}. \tag{12}$$

In n dimensions, the optimal h depends on the sample size N as

$$h \propto \frac{1}{N^{1/(n+4)}}. \tag{13}$$

Bias as model (mis)fit

The qualitative meaning of bias we will use has to do with the ability of the model class $\mathcal{F}$ to fit the data $\mathcal{D}_N$.

- We measure the misfit by the loss L associated with the task, i.e $\hat{L}(\hat{f}_{\mathcal{D}_N}, \mathcal{D}_N)$
- **Bias**$(\mathcal{F})$= $E_{P(X,Y)^n}[\hat{L}(\hat{f}_{\mathcal{D}_N}, \mathcal{D}_N)]$ (hence, bias is expected empirical loss).

- Richer model classes have less bias

$$\mathcal{F} \subset \mathcal{F}' \quad \text{then bias}(\mathcal{F}) \geq \text{bias}(\mathcal{F}')$$

- Larger data are harder to fit (hence more bias on average)[3]

[3] Not trivial, to find a reference.

Sampling variance

- Intuition: if we draw two different data sets $\mathcal{D}, \mathcal{D}' \sim P_{XY}$ (from the same distribution) we will obtain different predictors f, f'. Variance measures how different the predictions of f, f' can be on average.
- **Variance** at $x = Var_{P^n_{XY}}(\hat{f}_{\mathcal{D}_N}(x))$, where the randomness is over the sample $\mathcal{D}_N$
- **Variance** associated with predictor class $\mathcal{F}$ is the expectation over P_X of the variance at x, i.e $E_{P_X}[Var_{P^n_{XY}}(\hat{f}_{\mathcal{D}_N}(x))]$

- Variance depends on n, $\mathcal{F}$, and the data distribution P_{XY} Exercise If $P_{Y|X}$ is deterministic for all x, does it mean that the variance is 0?
- Richer model classes are subject to more variance

$$\mathcal{F} \subset \mathcal{F}' \quad \text{then} Var(\mathcal{F}) \leq Var(\mathcal{F}') \quad \text{for any } f^*$$

Variance, bias and model complexity

- Synonyms: rich class = complex model = flexible model = high modeling power = many degrees of freedom = many parameters

- Evaluating the model complexity[4]/number of free parameters of a model class $\mathcal{F}$ is usually a difficult problem!

  Non-parametric models # parameters depends on P_{XY}, smoothing parameter and n
  Parametric models # parameters NOT always equal to the number of parameters of f!

Example the classifier $f(x) = \text{sgn}(\alpha x), x, \alpha \in \mathbb{R}$ depends on one parameter α but has ∞ degrees of freedom[5]!

Example the linear classifier and regressor on $\mathbb{R}^d$ has (no more than) $n + 1$ degrees of freedom

Example the complexity of a two layer neural net with m fixed is not known (but there are approximation results); the number of weights in f is obviously $(m + 1)(n + 1) + 1$

Example For K-NN, the variance increases when K decreases

Example For pruned Decision Tree, the variance increases whith the number of levels

- The variance of a predictor increases with the complexity of $\mathcal{F}$.
- But complexity is the opposite of bias, so bias decrease with the complexity of $\mathcal{F}$
- This is known as the **Bias-Variance tradeoff**

[4] There are several definitions of model complexity, but this holds for all definitions I know

[5] See VC-dimension later

The Bias-Variance tradeoff

Wanted property (for an $\mathcal{F}$)	unwanted consequence of $\mathcal{F}$ not satisfying this property	what to do
to fit $\mathcal{D}$ well	Bias	increase complexity
to be robust to sampling noise	Variance	decrease complexity

The **bias-variance tradeoff** is the observation that the better a predictor class $\mathcal{F}$ is able to fit any given sample, the more sensitive the selected f will be to sampling noise.
In this course we will learn some ways of balancing these desired properties (or these undesired consequences).

Examples, examples. . .

Example (K-nearest neighbor classifiers)

The 1-NN can fit any data set perfectly (every data point is it's own nearest neighbor). But for $K > 1$, the K-NN may not be able to reproduce any pattern of ± 1 in the labels. Hence its bias is larger than the bias of the 1-NN classifier. With the variance, the opposite happens: as K the number of neighbors increases, the decision regions of the K-NN classifier become more stable to the random sampling effects. Thus, the variance decreases with K.

Example (Linear vs quadratic vs cubic . . . predictors)

The quadratic functions include all linear functions, the cubics include all quadratics, and so on. Linear classifiers will have more bias (less flexibility) than quadratic classifiers. On the other hand, the variance of the linear classifier will be lower than that of the quadratic. The case of regression is even more straightforward: if we fit the data with a higher degree polynomial, the fit will be more accurate, but the variation of the polynomial $f(x)$ for x values not in the training set will be higher too.

Example (Kernel regression)

Examples, examples... (2)

The bias-variance tradeoff can be observed on a continuous range for **kernel regression**. When the kernel width h is near 0, $f(x)$ from Lecture 1, equation (25) will fit the data in the training set exactly [Exercise: prove this], but will have high variance. When h is large, $f(x^i)$ will be smoothed between x^i and the other data points nearby, so it may be some distance from y^i. However, precisely because $f(x)$ is supported by a larger neighborhood, it will have low variance. [Exercise: find some intuitive explanations for why this is true] Hence, the smoothness parameter h controls the trade-off between bias and variance.

Example (Regularization)

The same can be observed if one considers equation (**??**). For $\lambda = 0$, one choses f that best fits the data (minimizes $\hat{L}$. For $\lambda \to \infty$, f is chosen to minimize the penalty J, disregarding the data completely. The latter case has 0 variance, but very large bias. Between these extreme cases, the parameter λ controls the amount in which we balance fitting the data (variance) with pulling f towards an a-priori "good" (bias).

Overfitting and Underfitting

- Bias and variance are properties of the model class $\mathcal{F}$ (sometimes toghether with the learning algorithm – more about this later). They are not properties of the parameters of f (e.g β), and not of a particular $f \in \mathcal{F}$.
- Variance decreases to 0 with n, but bias may not. This implies that for larger sample sizes n, the trade-off between variance and bias changes, and typically the "best" trade-off, aka the best model, will have larger complexity.
- **Overfitting**= is the situation of small bias and too much variance (i.e. $\mathcal{F}$ is too complex). In practice, if a learned predictor f has low $\hat{L}(f)$ but significantly higher $L(f)$, we say that the model has *overfit* the data $\mathcal{D}$. (Of course we cannot know $L(f)$ directly, and a significant amount of work in statistics is dedicated to predicting $L(f)$ for the purpose of chosing the best model.)
- **Underfitting**=bias is too high, or the model is too simple (a.k.a has too few degrees of freedom). [Exercise: what do you expect to see w.r.t. $\hat{L}(f)$ v.s. $L(f)$ for an underfitted model?]

Complexity, even though there are variations in its definition, and although it is not known exactly for most model classes, is at the core of learning theory, the part of statistical theory that gives provable results about the expected loss of a predictor.