

# Contents

|          |                                                     |             |
|----------|-----------------------------------------------------|-------------|
|          | <b>Preface</b> .....                                | <b>xi</b>   |
|          | <b>Acknowledgments</b> .....                        | <b>xiii</b> |
| <b>1</b> | <b>Automating Inquiry</b> .....                     | <b>1</b>    |
| 1.1      | A Thought Experiment .....                          | 1           |
| 1.2      | Active Learning .....                               | 3           |
| 1.3      | Scenarios for Active Learning .....                 | 5           |
| <b>2</b> | <b>Uncertainty Sampling</b> .....                   | <b>11</b>   |
| 2.1      | Pushing the Boundaries .....                        | 11          |
| 2.2      | An Example .....                                    | 12          |
| 2.3      | Measures of Uncertainty .....                       | 13          |
| 2.4      | Beyond Classification .....                         | 16          |
| 2.5      | Discussion .....                                    | 18          |
| <b>3</b> | <b>Searching Through the Hypothesis Space</b> ..... | <b>21</b>   |
| 3.1      | The Version Space .....                             | 21          |
| 3.2      | Uncertainty Sampling as Version Space Search .....  | 22          |
| 3.3      | Query by Disagreement .....                         | 24          |
| 3.4      | Query by Committee .....                            | 28          |
| 3.5      | Discussion .....                                    | 32          |
| <b>4</b> | <b>Minimizing Expected Error and Variance</b> ..... | <b>37</b>   |
| 4.1      | Expected Error Reduction .....                      | 37          |
| 4.2      | Variance Reduction .....                            | 40          |
| 4.3      | Batch Queries and Submodularity .....               | 44          |
| 4.4      | Discussion .....                                    | 46          |
| <b>5</b> | <b>Exploiting Structure in Data</b> .....           | <b>47</b>   |
| 5.1      | Density-Weighted Methods .....                      | 47          |
| 5.2      | Cluster-Based Active Learning .....                 | 49          |

|          |                                              |           |
|----------|----------------------------------------------|-----------|
| 5.3      | Active + Semi-Supervised Learning .....      | 53        |
| 5.4      | Discussion .....                             | 54        |
| <b>6</b> | <b>Theory .....</b>                          | <b>55</b> |
| 6.1      | A Unified View .....                         | 55        |
| 6.2      | A PAC Bound for Active Learning .....        | 57        |
| 6.3      | Discussion .....                             | 61        |
| <b>7</b> | <b>Practical Considerations .....</b>        | <b>63</b> |
| 7.1      | Which Algorithm is Best? .....               | 63        |
| 7.2      | Real Labeling Costs .....                    | 65        |
| 7.3      | Alternative Query Types .....                | 68        |
| 7.4      | Skewed Label Distributions .....             | 72        |
| 7.5      | Unreliable Oracles .....                     | 73        |
| 7.6      | Multi-Task Active Learning .....             | 74        |
| 7.7      | Data Reuse and the Unknown Model Class ..... | 76        |
| 7.8      | Stopping Criteria .....                      | 77        |
| <b>A</b> | <b>Nomenclature Reference .....</b>          | <b>79</b> |
|          | <b>Bibliography .....</b>                    | <b>81</b> |
|          | <b>Author's Biography .....</b>              | <b>97</b> |
|          | <b>Index .....</b>                           | <b>99</b> |